

CONFERENCE PROCEEDINGS

ICBO

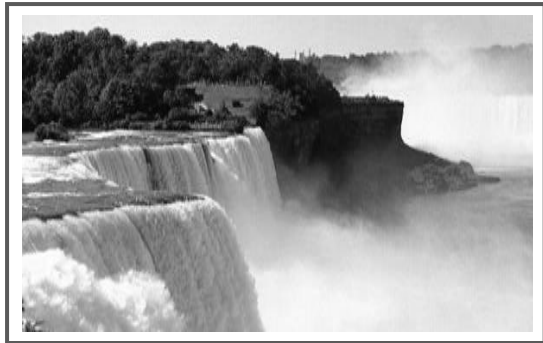
International Conference
on Biomedical Ontology

July 28-30, 2011
Buffalo, New York, USA



SPONSORED BY
University at Buffalo College of Arts and Sciences
National Center for Ontological Research
National Center for Biomedical Ontology
with support from
National Human Genome Research Institute

CONFERENCE PROCEEDINGS



ICBO

International Conference on Biomedical Ontology

July 28-30, 2011
Buffalo, New York, USA

Sponsored by:



University at Buffalo College of Arts and Sciences, National Center for Ontological Research, National Center for Biomedical Ontology, Computer Task Group, Cognigen Corporation, IMIA WG6 and the Protein Ontology

We would also like to acknowledge generous support from the National Human Genome Research Institute (NHGRI)

Funding for this conference was made possible in part by the United States National Institutes of Health (NIH) through the NIH Roadmap for Medical Research, Grant 1 U54 HG004028, and by 1 R13 HG006231-01 from the National Human Genome Research Institute (NHGRI). The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services or the National Institutes of Health; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government. Information on the National Centers for Biomedical Computing can be found at <http://nihroadmap.nih.gov/bioinformatics>.

PREFACE

In the two years that have elapsed since the first International Conference on Biomedical Ontology, held in Buffalo in 2009, the number of applications of ontologies in biomedicine has greatly increased, and so also has the range and functionality of associated software tools. The broad scope of current biomedical ontology research is well illustrated by the present volume, which includes contributions on ontologies at all scales, from genes, proteins and cells to mammalian anatomy, from biomedical investigations to the Electronic Health Record, and from adverse event reporting to the modeling of human physiology.

Like its predecessor, the second ICBO conference, also held in Buffalo, is based on the recognition that these different ontologies need to work together if we are to maximize the degree to which they can serve the needs of researchers and clinicians in supporting the integration of diverse bodies of data. The conference accordingly brings together scientists and informaticians developing and using ontologies across the entire spectrum of biology and clinical and translational medicine.

We are pleased to acknowledge the support of many individuals and institutions in the making of this conference. First, the NIH National Human Genome Research Institute provided a conference grant (R13HG006231-01), which funded a number of generous fellowships to early career participants, and also funds for the National Center for Biomedical Ontology (NCBO), which provided valuable scientific and organizational assistance, and sponsored a panel on the topic of ontology technology in support of clinical and translational Science. Second, the Protein Ontology Consortium, who sponsored a panel on proteins and diseases. Third, the members of the Organizing and Program Committees, and especially Olivier Bodenreider, for indispensable support in crucial phases of the conference preparation. Fourth, Computer Task Group and Cognigen Corporation, for their sponsorship of conference receptions. Fifth, the University at Buffalo for providing logistical services. And finally, Sandra G. Smith, for exercising her conferencing organizing skills in ways which go far beyond the call of duty.

Barry Smith
Conference Chair
Buffalo, New York, USA
July 2011

ICBO ORGANIZING COMMITTEE

Conference Chair: Barry Smith, University at Buffalo, Buffalo, NY, USA

Judith A. Blake, The Jackson Laboratory, Bar Harbor, ME, USA

Suzanna E. Lewis, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Mark A. Musen, Stanford University, Stanford, CA, USA

Alan Ruttenberg, University at Buffalo, Buffalo, NY, USA

Susanna-Assunta Sansone, Oxford University, Oxford, UK

Dagobert Soergel, University at Buffalo, Buffalo, NY, USA

Christian J. Stoeckert Jr., University of Pennsylvania, Philadelphia, PA USA

ICBO PROGRAM COMMITTEE

Program Committee Chair: Alan Ruttenberg, University at Buffalo, Buffalo, NY USA

Co-Chair: Olivier Bodenreider, NIH National Library of Medicine, Bethesda, MD, USA

Co-Chair: Maryann E. Martone, University of California at San Diego, San Diego, CA, USA

Workshops and Tutorials Chair: Stefan Schulz, Medical University of Graz, Austria

Doctoral and Postdoctoral Consortium Chair: Albert Goldfain, University at Buffalo, Buffalo, NY, USA

Software Demonstrations Chair: Patricia L. Whetzel, Stanford University, Stanford, CA, USA

Colin Batchelor, Royal Society of Chemistry, Cambridge, UK

Thomas Beale, Ocean Informatics

Judith A. Blake, The Jackson Laboratory, Bar Harbor, ME, USA

Sebastian Brandt, University of Manchester, Manchester, UK

Werner Ceusters, University at Buffalo, Buffalo, NY, USA

Rex Chisholm, Northwestern University, Chicago, IL, USA

Jim Cimino, National Institutes of Health, Bethesda, MD, USA

Mélanie Courtot, British Columbia Cancer Research Centre, Vancouver, BC, Canada

Lindsay G. Cowell, University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA

Alexander D. Diehl, University at Buffalo, Buffalo, NY, USA

Michel Dumontier, Carleton University, Ottawa, ON, Canada

Louis J. Goldberg, University at Buffalo, Buffalo, NY, USA

Janna Hastings, European Bioinformatics Institute, Hinxton, UK

Pascal Hitzler, Wright State University, Dayton, OH, USA

Robert Hoehndorf, University of Cambridge, Cambridge, UK

Larry Hunter, University of Colorado, Denver, CO, USA

Marijke Keet, School of Computer Science, University of KwaZulu-Natal, Durban, South Africa

Jobst Landgrebe, Cognitekt UG, Cologne, Germany

Suzanna E. Lewis, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Philip Lord, Newcastle University, Newcastle upon Tyne, UK

Jose L. V. Mejino, Jr., University of Washington, Seattle, WA, USA

Alexa McCray, Harvard Medical School, Boston, MA, USA

Christopher J. Mungall, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Mark A. Musen, Stanford University, Stanford, CA, USA

Darren Natale, Georgetown University, Washington, DC, USA

Chimezie Ogbuji, Case Western Reserve University, Cleveland, OH, USA

Helen Parkinson, European Bioinformatics Institute, Hinxton, UK

Bjoern Peters, La Jolla Institute for Allergy and Immunology, San Diego, CA, USA

Daniel Rubin, Stanford University, Stanford, CA, USA

Peter Robinson, Charité Hospital, Berlin

Susanna-Assunta Sansone, Oxford University, Oxford, UK

Ulrike Sattler, University of Manchester, Manchester, UK

Richard Scheuermann, University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA

Barry Smith, University at Buffalo, Buffalo, NY, USA

Dagobert Soergel, University at Buffalo, Buffalo, NY, USA

Robert Stevens, University of Manchester, Manchester, UK

Ida Sim, University of California at San Francisco, San Francisco, CA, USA

Harold Solbrig, Mayo Clinic, Rochester, MN, USA

Kent Spackman, International Health Terminology Standards Development Organization

Christian J. Stoeckert Jr., University of Pennsylvania, Philadelphia, PA USA

Cathy Wu, Georgetown University, Washington, DC, USA; University of Delaware, Newark, DE, USA

TABLE OF CONTENTS

Preface	iii
ICBO Committees	v
ICBO Presentations	1
Where GO is Going and What it Means for Ontology Extension <i>Catia Pesquita, Francisco M. Couto</i>	3
Revising the Cell Ontology <i>Terrence F. Meehan, Christopher J. Mungall, Alexander D. Diehl</i>	11
Rapid Development of an Ontology of Coriell Cell Lines <i>Chao Pang, Tomasz Adamusiak, Helen Parkinson, James Malone</i>	19
Cell Line Ontology: Redesigning the Cell Line Knowledgebase to Aid Integrative Translational Informatics <i>Sirarat Samtivijai, Zuoshuang Xiang, Terrence F. Meehan, Alexander D. Diehl, Uma Vempati, Stephan Schurer, Chao Pang, James Malone, Helen Parkinson, Brian D. Athey, Yongqun He</i>	25
Towards a Semantic Web Application: Ontology-Driven Ortholog Clustering Analysis <i>Yu Lin, Zuoshuang Xiang, Yongqun He</i>	33
HeartCyc: A Cardiac Cycle Process Ontology Based in the Ontology of Physics for Biology <i>Daniel L. Cook, Michal Galdzicki, Maxwell L. Neal, Jose L.V. Mejino, John H. Gennari</i>	41
Composite Annotation for Heart Development <i>Tariq Abdulla, Ryan Imms, Jean-Marc Schleich, Ron Summers</i>	47
Ontology-Based Analysis of Event-Related Potentials <i>Gwen Frishkoff, Robert Frank, Paea LePend</i>	55
River Flow Model of Diseases <i>Riichiro Mizoguchi, Kouji Kozaki, Hiroko Kou, Yuki Yamagata, Takeshi Imai, Kayo Waki, Kazuhiko Ohe</i>	63
Dispositions and Processes in the Emotion Ontology <i>Janna Hastings, Werner Ceusters, Barry Smith, Kevin Mulligan</i>	71
An Advanced Strategy for Integration of Biological Measurement Data <i>Hiroshi Masuya, Georgios V. Gkoutos, Nobuhiko Tanaka, Kazunori Waki, Yoshihiro Okuda, Tatsuya Kushida, Norio Kobayashi, Koji Doi, Kouji Kozaki, Robert Hoehndorf, Shigeharu Wakana, Tetsuro Toyoda, Riichiro Mizoguchi</i>	79
Annotating Experimental Records Using Ontologies <i>Alexander García, Olga Giraldo, Jael García</i>	87
Ontology Driven Data Collection for EuPathDB <i>Jie Zheng, Omar S. Harb, Christian J. Stoeckert Jr.</i>	95
Developing an Application Ontology for Biomedical Resource Annotation and Retrieval: Challenges and Lessons Learned <i>Carlo Torniai, Matt Brush, Nicole Vasilevsky, Erik Segerdell, Melanie Wilson, Tenille Johnson, Karen Corday, Chris Shaffer, Melissa Haendel,</i>	101
Mapping Composition for Matching Large Life Science Ontologies <i>Anika Gross, Michael Hartung, Toralf Kirsten, Erhard Rahm</i>	109
Generic Semantic Relatedness Measure for Biomedical Ontologies <i>João D. Ferreira, Francisco M. Couto</i>	117

Aligning the Parasite Experiment Ontology and the Ontology for Biomedical Investigations Using AgreementMaker	125
<i>Valerie Cross, Cosmin Stroe, Xueheng Hu, Pramit Silwal, Maryam Panahiazar, Isabel F. Cruz, Priti Parikh, Amit Sheth</i>	
A Case Study of ICD-11 Anatomy Value Set Extraction from SNOMED CT	133
<i>Guoqian Jiang, Harold R. Solbrig, Robert J.G. Chalmers, Kent Spackman, Alan L. Rector, Christopher G. Chute</i>	
The HL7 Approach to Semantic Interoperability	139
<i>Jobst Landgrebe, Barry Smith</i>	
Representing the Reality Underlying Demographic Data	147
<i>William R. Hogan, Swetha Garimalla, Shariq A. Tariq</i>	
Information Models and Ontologies for Representing the Electronic Health Record	153
<i>Daniel Karlsson, Martin Berzell, Stefan Schulz</i>	
Ontology-Based Mammography Annotation and Similar Mass Retrieval with SQWRL	159
<i>Hakan Bulu, Adil Alpkocak, Pinar Balci</i>	
The CHRONIOUS Ontology Suite: Methodology and Design Principles	167
<i>Luc Schneider, Mathias Brochhausen</i>	
Applying Rigidity to Standardizing OBO Foundry Candidate Ontologies	175
<i>Patrice Seyed, Stuart C. Shapiro</i>	
Higgs Bosons, Mars Missions, and Unicorn Delusions: How to Deal with Terms of Dubious Reference in Scientific Ontologies	183
<i>Stefan Schulz, Mathias Brochhausen, Robert Hoehndorf</i>	
Towards an Ontology for Conceptual Modeling	191
<i>James P. McCusker, Joanne Luciano, Deborah L. McGuinness</i>	
What's in an 'is about' Link? Chemical Diagrams and the Information Artifact Ontology	201
<i>Janna Hastings, Colin Batchelor, Fabian Neuhaus, Christoph Steinbeck</i>	
Bioassay Ontology to Describe High-Throughput Screening Assays and their Results	209
<i>Uma Vempati, Ubbo Visser, Saminda Abeyruwan, Kunie Sakurai, Magdalena Przydzial, Caty Chung, Robin Smith, Amar Koleti, Christopher Mader, Vance Lemmon, Stephan Schürer</i>	
A Framework Ontology for Computer-Based Patient Record Systems	217
<i>Chimezie Ogbuji</i>	
ICBO Posters	225
The Blood Ontology: An Ontology in the Domain of Hematology	227
<i>Mauricio Barcellos Almeida, Anna Barbara de Freitas Carneiro Proietti, Jiye Ai, Barry Smith</i>	
Employing Reasoning within the Phenoscape Knowledgebase	230
<i>James P. Balhoff, Wasila M. Dahdul, Hilmar Lapp, Peter E. Midford, Todd J. Vision, Monte Westerfield, Paula M. Mabee</i>	
Waiting for a Robust Disease Ontology: A Merger of OMIM and MeSH as a Practical Interim Solution	231
<i>Susan M. Bello, Allan Peter Davis, Thomas C. Wieggers, Mary E. Dolan, Cynthia Smith, Joel Richardson, Judith Blake, Carolyn Mattingly, Janan T. Eppig</i>	
Developing a Reagent Application Ontology within the OBO Foundry Framework	234
<i>Matthew H. Brush, Nicole Vasilevsky, Carlo Torniai, Tenille Johnson, Chris Shaffer, Melissa A. Haendel</i>	

The Ontology of Microbial Phenotypes (OMP): A Precomposed Ontology Based on Cross Products from Multiple External Ontologies that is Used for Guiding Microbial Phenotype Annotation	237
<i>Marcus Chibucos, Adrienne Zweifel, Deborah Siegele, Peter Uetz, Michelle Giglio, James Hu</i>	
Towards an Adverse Event Reporting Ontology	240
<i>Mélanie Courtot, Ryan R. Brinkman</i>	
OntoOrpha : An Ontology to Support Edition and Audit of Knowledge of Rare Diseases in ORPHANET	241
<i>Ferdinand Dhombres, Pierre-Yves Vandenbussche, Ana Rath, Marc Hanauer, Annie Olry, Bruno Urbero, Rémy Choquet, Jean Charlet</i>	
An Ontology for Gastrointestinal Endoscopy	244
<i>Shahim Essaid</i>	
Enriching the Ontology for Biomedical Investigations (OBI) to Improve its Suitability for Web Service Annotations	246
<i>Chaitanya Guttula, Alok Dhamanaskar, Rui Wang, John A. Miller, Jessica C. Kissinger, Jie Zheng, Christian J. Stoeckert, Jr.</i>	
Recent Developments in the ChEBI Ontology	249
<i>Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Kenneth Haug, Zara Josephs, Gareth Owen, Steve Turner, Christoph Steinbeck</i>	
Representing Local Identifiers in a Referent-Tracking System	252
<i>William R. Hogan, Swetha Garimalla, Shariq A. Tariq, Werner Ceusters</i>	
owl_cpp, a C++ Library for Working with OWL Ontologies	255
<i>Mikhail K. Levin, Alan Ruttenberg, Anna Maria Masci, Lindsay G. Cowell</i>	
Towards Ontologies for 'Textbook' of the Future in Obstetrics and Gynecology (OB/Gyn)	258
<i>Yu Lin, Chris Chapman, Erin D. Doelling, Maya Hammoud</i>	
Gene Ontology Signatures for Immune Cell-Types Inferred by Gene Expression Analysis	259
<i>Terrence F. Meehan, Christopher J. Mungall, Judith A. Blake, Alexander D. Diehl</i>	
Aligning Research Resource and Researcher Representation: The eagle-i and VIVO Use Case.	260
<i>Stella Mitchell, Carlo Torniai, Brian Lowe, Jon Corson-Rikert, Melanie Wilson, Mansoor Ahmed, Shanshan Chen, Ying Ding, Nicholas Rejack, Melissa Haendel, the eagle-i Consortium, the VIVO Collaboration</i>	
An Ontology-Based Approach to Linking Model Organisms and Resources to Human Diseases	263
<i>Chris J. Mungall, David Anderson, Anita Bandrowski, Brian Canada, Andrew Chatr-Aryamontri, Keith Cheng, P. Michael Conn, Kara Dolinski, Mark Ellisman, Janan Eppig, Jeffrey S. Grethe, Joseph Kemnitz, Shawn Iadonato, Stephen D. Larson, Charles Magness, Maryann E. Martone, Mike Tyers, Carlo Torniai, Olga Troyanskaya, Judith Turner, Monte Westerfield, Melissa A. Haendel</i>	
Using RxNorm to Extract Medication Data from Electronic Health Records in the Rochester Epidemiology Project	266
<i>Jyotishman Pathak, Shawn P. Murphy, Brian N. Willaert, Hilal M. Kremers, Christopher G. Chute, Barbara P. Yawn, Walter A. Rocca</i>	
Towards Desiderata for Provenance Ontologies in Biomedicine	269
<i>Satya S. Sahoo</i>	
Modeling Issues and Solutions: Building a Taxonomy from a Biology Textbook	273
<i>Patrice Seyed, John Pacheco, Andrew Goldenkranz, Vinay Chaudhri</i>	
Biomedical Analyses: OWL Model Based Edition	276
<i>Pierre-Yves Vandenbussche, Ferdinand Dhombres, Sylvie Cormont, Jean Charlet, Eric Lepage</i>	
Ontobee: A Linked Data Server and Browser for Ontology Terms	279
<i>Zuoshuang Xiang, Chris Mungall, Alan Ruttenberg, Yongqun He</i>	
ICBO Software Demonstrations.	283

Protein-Centric Connection of Biomedical Knowledge: Protein Ontology (PRO) Research and Annotation Tools . .	285
<i>Cecilia N. Arighi, Darren A. Natale, Judith A. Blake, Carol J. Bult, Michael Caudy, Alexander D. Diehl, Harold J. Drabkin, Peter D'Eustachio, Alexei Evsikov, Hongzhan Huang, Natalia V. Roberts, Alan Ruttenberg, Barry Smith, Jian Zhang, Cathy H. Wu</i>	
SHIRAZ and CABERNET: Leveraging Automation, Crowdsourcing, and Ontologies to Improve the Accuracy and Throughput of Zebrafish Histological Phenotype Annotations.	288
<i>Brian Canada, Georgia Thomas, John Schleicher, James Z. Wang, Keith C. Cheng</i>	
Biomedical Ontology Matching Using the AgreementMaker System	290
<i>Isabel F. Cruz, Cosmin Stroe, Catia Pesquita, Francisco M. Couto, Valerie Cross</i>	
Bioportal: Ontologies and Integrated Data Resources at the Click of a Mouse.	292
<i>Patricia L. Whetzel, Natasha Noy, Nigam Shah, Paul Alexander, Michael Dorf, Ray Ferguson, Margaret-Anne Storey, Barry Smith, Chris Chute, Mark Musen</i>	
Populous: A Tool for Populating an Ontology	294
<i>Simon Jupp, Matthew Horridge, Luigi Iannone, Julie Klein, Stuart Owen, Joost Schanstra, Katy Wolstencroft, Robert Stevens</i>	
The Vitro Integrated Ontology Editor and Semantic Web Application	296
<i>Brian Lowe, Brian Caruso, Nick Cappadona, Miles Worthington, Stella Mitchell, Jon Corson-Rikert, and VIVO Collaboration</i>	
The BioPortal Import Plugin for Protégé	298
<i>Jithun Nair, Tania Tudorache, Trish Whetzel, Natalya Noy, Mark Musen</i>	
The Biomedical Ontology Applications (BOA) Framework	300
<i>Bruno Tavares, Hugo P. Bastos, Daniel Faria, João D. Ferreira, Tiago Grego, Catia Pesquita, Francisco M. Couto</i>	
The NCBO Annotator: Ontology-Based Annotation as a Web Service	302
<i>Patricia L. Whetzel, Clement Jonquet, Cherie Youn, Michael Dorf, Ray Ferguson, Mark Musen, Nigam Shah</i>	
OntoFox and its Application in Development of Brucellosis Ontology	304
<i>Zuoshuang Xiang, Yu Lin, Yongqun He</i>	

Pre-ICBO Workshops

Tuesday, July 26, 2011

Representing Adverse Events.	307
<i>AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events 309</i>	
<i>Yongqun He, Zuoshuang Xiang, Sirarat Santivijai, Luca Toldo, Werner Ceusters</i>	
Reporting Adverse Events: Basis for a Common Representation	316
<i>Mélanie Courtot, Ryan R. Brinkman, Alan Ruttenberg</i>	
Adverse Events from Clinical Studies in Pharmaceutical Research and Development	324
<i>Pia Emilsson, Kerstin Forsberg</i>	
Classifying Adverse Events from Clinical Trials	326
<i>Bernard LaSalle, Richard Bradshaw</i>	
An Ontological Representation of Adverse Drug Events	329
<i>Guoqian Jiang, Jon D. Duke, Jyotishman Pathak, Christopher G. Chute</i>	
Toward Answering Time-Related Questions from Adverse Event Reports Using Ontology-Based Approaches . .	332
<i>Cui Tao, Guoqian Jiang, Kim Clark, Deepak Sharma, Christopher G. Chute</i>	
Supporting Medical Device Adverse Event Analysis in an Interoperable Clinical Environment: Design of a Data Logging and Playback System	335
<i>David Arney, Sandy Weininger, Susan F. Whitehead, Julian M. Goldman</i>	

Using Failure Modes, Mechanisms, and Effects Analysis in Medical Device Adverse Event Investigations	340
<i>Shunfeng Cheng, Diganta Das, Michael Pecht</i>	

Working with Multiple Biomedical Ontologies 347

NIFSTD and NeuroLex: A Comprehensive Neuroscience Ontology Development Based on Multiple Biomedical Ontologies and Community Involvement	349
<i>Fahim T. Imam, Stephen D. Larson, Jeffrey S. Grethe, Amarnath Gupta, Anita Bandrowski, Maryann E. Martone</i>	
Use of Multiple Ontologies to Characterize the Bioactivity of Small Molecules	357
<i>Ying Yan, Janna Hastings, Jee-Hyub Kim, Stefan Schulz, Christoph Steinbeck, Dietrich Rebholz-Schuhmann</i>	
A Meta-Data Approach to Querying Multiple Biomedical Ontologies	364
<i>Ravi Palla, Dan Tecuci, Vinay Shet, Mathaeus Dejori</i>	
Multiple Ontologies in Healthcare Information Technology: Motivations and Recommendation for Ontology Mapping and Alignment	367
<i>Colin Puri, Karthik Gomadam, Prateek Jain, Peter Z. Yeh, Kunal Verma</i>	
Modularization for the Cell Ontology	370
<i>Christopher J. Mungall, Melissa Haendel, Amelia Ireland, Shahid Manzoor, Terry Meehan, David Osumi-Sutherland, Carlo Torniai, Alexander D. Diehl</i>	
Building the OBO Foundry – One Policy at a Time	377
<i>Mélanie Courtot, Chris Mungall, Ryan R. Brinkman, Alan Ruttenberg</i>	
Towards a Body Fluids Ontology: A Unified Application Ontology for Basic and Translational Science	381
<i>Jiye Ai, Mauricio Barcellos Almeida, André Queiroz de Andrade, Alan Ruttenberg, David Tai Wai Wong, Barry Smith</i>	
Connecting Ontologies for the Representation of Biological Pathways	387
<i>Anna Maria Masci, Mikhail Levin, Alan Ruttenberg, Lindsay G. Cowell</i>	
Bridging Multiple Ontologies: Representation of the Liver Immune Response	393
<i>Anna Maria Masci, Jeffrey Roach, Bernard de Bono, Pierre Grenon, Lindsay Cowell</i>	
Hyperontology for the Biomedical Ontologist: A Sketch and Some Examples	399
<i>Oliver Kutz, Till Mosskowsky, Janna Hastings, Alexander Garcia Castro, Aleksandra Sojic</i>	

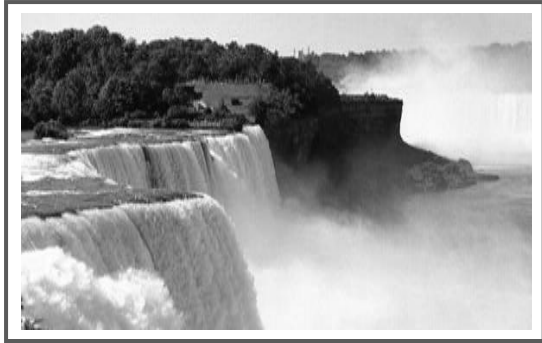
Wednesday, July 27, 2011

Facilitating Anatomy Ontology Interoperability 409

Mouse Anatomy Ontologies and GXD	411
<i>Terry F. Hayamizu, Martin Ringwald</i>	
FUNCARO: A Functional Extension to CARO	412
<i>David Osumi-Sutherland</i>	
The Vertebrate Bridging Ontology (VBO)	416
<i>Ravensara Travillian, James Malone, Chao Pang, John Hancock, Peter W.H. Holland, Paul Schofield, Helen Parkinson</i>	
Multi-Species Anatomy Ontology Development Requires a Pluralist Approach to Label-Class Mapping	422
<i>István Mikó, Matthew J. Yoder, Matthew A. Bertone, Katja C. Seltsmann, Andrew R. Deans</i>	
CARO 2.0	425
<i>David Osumi-Sutherland</i>	
Integrating Anatomy and Phenotype Ontologies with Taxonomic Hierarchies	426
<i>James P. Balhoff, Peter E. Midford, Hilmar Lapp</i>	
Phenoscape: Use Cases and Anatomy Ontology Requirements for Linking Evolutionary and Model Organism Phenotypes	428
<i>Wasila Dahdul, James Balhoff, Hilmar Lapp, Peter Midford, Todd Vision, Monte Westerfield, Paula Mabee</i>	

Ontologies in the Fish Tank: Using the Zebrafish Anatomy Ontology with Other OBO Ontologies to Annotate Expression and Phenotype.	430
<i>Yvonne Bradford, Ceri Van Slyke</i>	
Doctoral and Post-Doctoral Consortium.	431
One Empirical Study, Three Tests of Translation, Many Questions on Biomedical Ontologies: Limited Contribution of MeSH Terms to Effective Literature Searches on ‘Health-Related Values’	433
<i>Mila Petrova</i>	
The Age of Data-Driven Medicine: Mining the Electronic Health Record	435
<i>Paea LePendou, Mark A. Musen, Nigam H. Shah</i>	
Quality of Care Domain Modeling in Cancer: A Semantic Approach	436
<i>Sina Madani, Dean F. Sittig, Parsa Mirhaji, Kim Dunn</i>	
Philosophy, Ontology, and Scientific Explanation.	438
<i>James A. Overton</i>	
Refining Ontology for Glucose Metabolism Disorders	440
<i>Yu Lin</i>	
A Case Study in Using ZFA and PATO for Describing Histological Phenotypes in the Larval Zebrafish.	441
<i>Brian Canada, Georgia Thomas, Timothy Cooper, Keith Cheng</i>	
Epistemological Issues in Information Organization Instruments: Ontologies and Health Information Models.	442
<i>André Queiroz Andrade, Mauricio Barcellos Almeida</i>	
The Foundational Model of Neuroanatomy Ontology: An Ontology Framework to Support Neuroanatomical Data Integration	444
<i>B. Nolan Nichols, Jose L.V. Mejino Jr., James F. Brinkley</i>	
Workflows and Framework for Nutritional/Metabolic Phenotype/Genotype and Foods-for-Health Knowledge Integration	446
<i>Matthew Lange, J. Bruce German, Jim Kaput</i>	
Applications for a Translational Biomedical Ontology Model	447
<i>Robert Yao, Graciela Gonzalez</i>	
Author Index	449

ICBO Presentations



ICBO

International Conference on Biomedical Ontology

July 28-30, 2011
Buffalo, New York, USA

Where GO is Going and What it Means for Ontology Extension

Catia Pesquita, Francisco M. Couto

Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa Campo Grande, Lisboa, Portugal

Abstract. Developing and maintaining a biomedical ontology is a time- and effort-consuming task, given the dynamic and expanding nature of biomedical knowledge. This is a relevant issue for very large ontologies which cover a broad domain, for smaller ontologies maintained by a small team and also for domains where being able to perform quick updates is critical (e.g. epidemiology).

The first step in the process of extending an ontology is identifying the areas of the ontology that need to be changed – change capturing. In this paper we propose that this process can be semi-automated by exploring ontology information. This would be a valuable tool to support ontology developers in ontology extension, easing their burden.

In order to accomplish this, we have developed a framework for analysing the extension of ontologies, to create a general panorama of ontology extension processes that can guide the development of change capturing techniques. We have applied it to the analysis of the extension of the Gene Ontology and uncovered some of the underlying tendencies in its extension. Building upon the results of this analysis and a set of guidelines for ontology change capturing, we then investigated the feasibility of predicting which classes of the ontology will be extended in a future version. Finally, we discuss the obtained results and identify the main challenges and future directions for the budding area of ontology extension prediction.

Keywords: Biomedical ontologies, ontology extension, ontology refinement, ontology enrichment, prediction of ontology extension

1 Introduction

The development of a biomedical ontology is a very demanding process that requires both expertise in the domain to model and in ontology design. It is also necessarily an iterative process [10] since biomedical knowledge is diverse, complex and continuously changing and growing. This process, usually named ontology evolution, requires large investments of both time and money with each new ontology version that is produced.

Ontology evolution can be defined as the process of modifying an ontology in response to a certain change in the domain or its conceptualization [6]. These include changes in the portion of the real world they model, the uncovering of information previously unavailable, a reassessment of the relevance of some element to the ontology or a need to correct previous mistakes [4]. In the last couple

of years, a generally agreed upon model for ontology evolution has emerged, which includes four distinct steps: (1) requesting the change, (2) planning the change, (3) implementing the change and (4) verification and validation (for a review see [8]). The changes made in the course of ontology evolution can be of three elementary types: addition, removal and modification [13]. We define ontology extension as the process of ontology evolution concerned with the addition of new elements. We consider ontology extension to encompass both ontology refinement (the addition of new classes to an ontology) and ontology enrichment (the addition of non-taxonomical relations or richer axioms).

Before these changes are actually performed, the need for the change must be identified. This is the first step in ontology evolution, the change capturing phase [12], and it can be based on explicit or implicit requirements [5]. Explicit requirements correspond to those made by the ontology

developers or to requests made by end-users. Implicit requirements correspond to those that can be uncovered by change discovery. Stojanovic et al. [13] list a series of guidelines for change capturing, organized into three types according to the kind of data they exploit, to which Castaño et al.[2] add a fourth:

structure-driven: which are derived from the structure of the ontology, e.g. ‘A class with a single subclass should be merged with its subclass’.

data-driven: which correspond to implicit changes in the domain and are discovered through the analysis of the instances belonging to the ontology, e.g. ‘A class with many instances is a candidate for being split into subclasses and its instances distributed among newly generated classes’.

usage-driven: which are deduced from the usage patterns of the ontology in the knowledge management system e.g. classes that have not been retrieved in a long time might be out of date.

discovery-driven: which is applied when a new instance cannot be described by the ontology classes, and new classes are identified using external resources.

These changes can in principle be semi-automatically discovered by analyzing the ontology data and its usage. This process can be cast in terms of a prediction of ontology extension and can represent a significant contribution to easing the burden of keeping an ontology up-to-date.

The main application of a methodology to predict ontology extension is to support manual or semi-automated ontology extension, since it can minimize the effort in collecting and crossing the information necessary to make the extension decisions. It can contribute to the evolution of larger ontologies by helping to pinpoint the areas that are in need of attention and also to smaller ontologies, where team size and resources may not be as large, by reducing the time and effort spent. It can also provide valuable assistance when there is an urgent need to extend an ontology to cover a new aspect, such as in the case of an epidemics, where the inclusion of new classes in a timely fashion can improve the performance of data analysis methods [11]. Furthermore, they can

be incorporated into automated and semi-automated ontology extension systems, which so far have not addressed this issue [7, 9, 14].

To guide the development of methods for automated prediction of ontology extension it is of interest to analyze the extension of ontologies. In [4], a methodology for calculating the improvements obtained in successive versions of biomedical ontologies based on the matches and mismatches between them is proposed. It has been used to calculate the degree of correctness of the Gene Ontology (GO) terminology and to forecast how this overall quality will improve [3]. In this detailed analysis of the evolution of GO 75% of the changes made to classes were identified as being insertions. However this study has a strong focus on error correction, which includes not only the addition of new elements, but also their removal: of the 17 parameters used, only two correspond to the absence of elements that should be included in the ontology. Furthermore, the method does not take into account the hierarchical level at which the error is made, nor does it consider GO annotations.

In this paper we present a preliminary framework for analysing the extension of an ontology, and apply it to the analysis of GO. We have chosen GO for our study because it presents a very interesting case: it is the most prominent bioontology, with widespread use and impact; it covers a considerable wide domain; it provides a corpus of annotations and it is updated on a frequent basis, thus supporting the investigation of its evolution through the analysis of different versions. Building upon the analysis of ontology extension, we then investigate the feasibility of predicting the evolution of GO. Since GO authors do not justify the addition of new elements, we based our prediction in a set of rules derived from the guidelines for ontology change capture. We were interested in evaluating the suitability of these guidelines as a support for ontology extension prediction. Finally, we discuss the novel area of ontology extension prediction, its challenges and role in the future of ontology development.

2 A Framework for Analyzing Ontology Extension

An analysis of ontology extension should by

definition, focus on both refinement and enrichment, and analyze several versions of the same ontology during a time period. The decision on the time period to analyze should be based on the age of the ontology as well as the availability and frequency of new version releases.

For analyzing ontology refinement we propose inspecting three key aspects:

1. depth of new classes, i.e. minimum distance to the root class over *is_a* and *part_of* relations.
2. number of new classes that are children of existing *vs.* newly added classes
3. number of new classes that are children of leaf classes

The first and third aspects capture the general direction of the refinement of the ontology, where additions at a greater depth and to leaf classes represent vertical extension whereas additions at middle depth and to non-leaf classes represent horizontal extension. These aspects are helpful to analyze the level of detail and coverage provided by the refinement. The second aspect is related to another interesting characteristic of refinement, whether new classes are inserted individually or whether as part of a new branch.

For analyzing ontology enrichment we propose investigating the following:

4. age and depth of the classes linked by the new relation (i.e. whether the relation is

established between old classes, between an old and a new class or between new classes)

This aspect is intended to capture first at what level of specificity do the enrichment events take place, and secondly if enrichment happens alongside refinement or if it succeeds it.

2.1 Analyzing the Gene Ontology Extension

Based on the aspects identified in the previous section, we have analyzed 12 versions of the Gene Ontology and its annotations equally spaced over a period of 6 years (2005-2010). At 6 month intervals, new classes represent about 5% of all classes in the ontology. In the context of GO, enrichment corresponds to the insertion of new non *is_a* relations between existing or newly inserted classes.

Figures 1, 2, 3 and 4 show the results of the analysis of each aspect. In all three hierarchies, the majority of new subclasses are added as children of non-leaf classes, resulting in a prevalence of horizontal extension. Also, the refinement of molecular function and cellular component occurs mostly via single insertions, whereas in the biological process groups of related classes are inserted together. Regarding enrichment, in biological process, a considerable portion of relations are established between two newly inserted classes, whereas in cellular component, the majority is made between an existing and a new class.

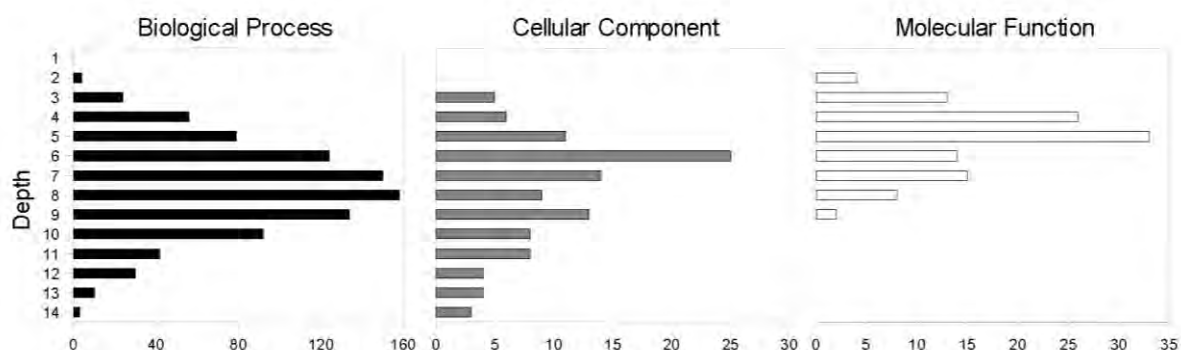


Figure 1. Average depth of new classes

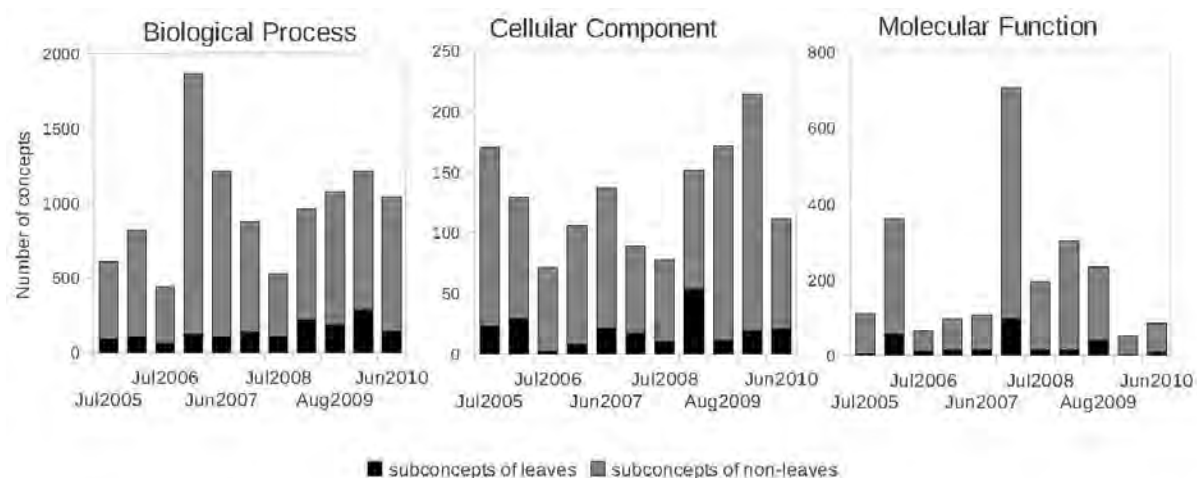


Figure 2. Ancestry of new classes (leaves or non-leaves) by ontology version

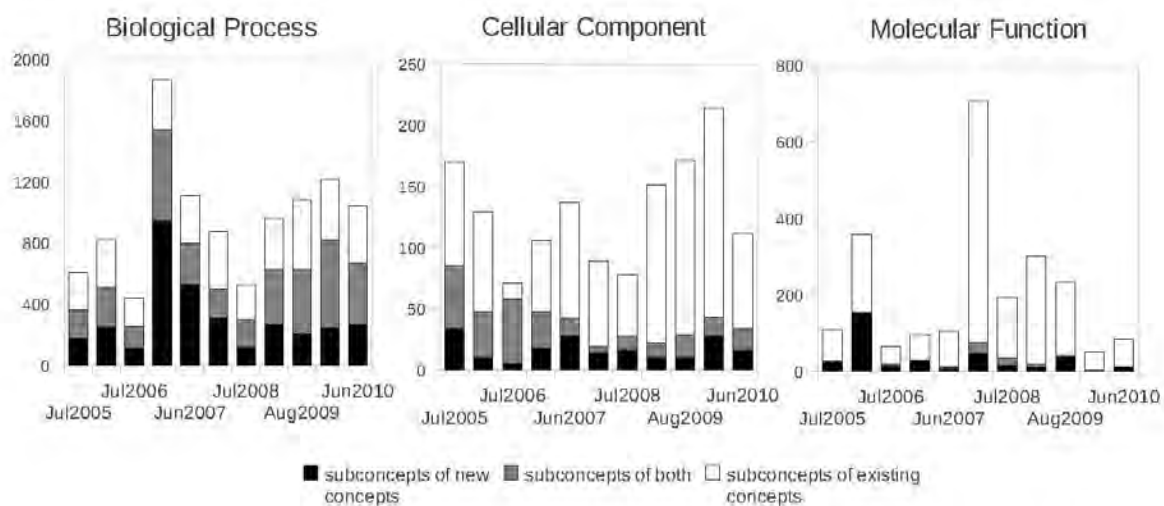


Figure 3. Ancestry of new classes (existing or new parents) by ontology version

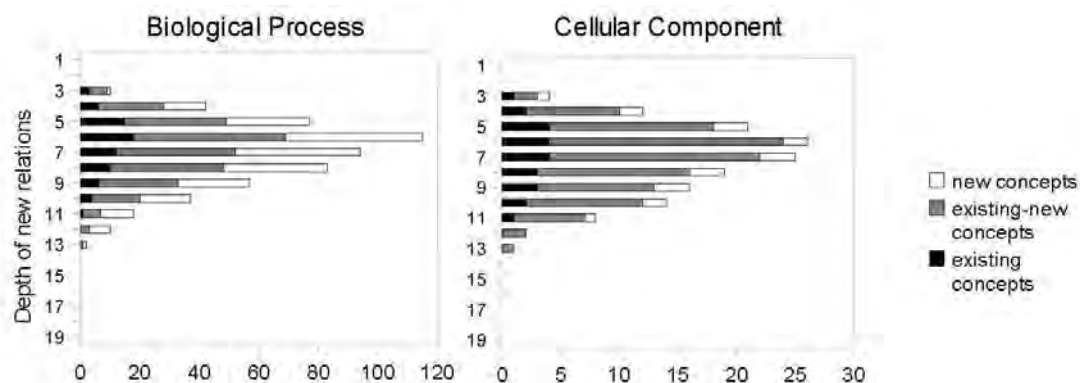


Figure 4. Depth and age of the classes in new enrichment relations. Molecular Function not shown, since it contains less than 10 non *is_a* relations.

3 Predicting Ontology Extension: A Rule-Based Approach

Adapting and extending the guidelines proposed by [13] following [10] that are concerned with ontology extension, we recognize two heuristics to identify potential ontology extensions and classify them according to the type of data they use:

1. structure-driven: If a class has fewer children than its siblings, it may be a candidate for extension
2. data-driven: A class with many instances is a candidate for being split into subclasses and its instances distributed among newly generated classes.

Following the above mentioned guidelines we have devised a set of rules to apply to the prediction of the extension of GO. The rules aim at finding a partition of the set of classes that best separates classes that will be refined in a future version from those that will not. Here, we assume that the latest ontology version is believed to be as correct and complete as possible [4]. We have three types of rules, one structure-based and two data-based. The structure-based rules are derived from guideline 1:

Rule 1: A class with at most $x\%$ fewer subclasses than its siblings is a candidate for refinement

with x taking four evenly spaced values between 25 and 100%. The data-based rules are derived from guideline 2 but distinguish between the set of all annotations and the set of manually curated ones:

Rule 2: A class with at least $x\%$ more annotations than its siblings is a candidate for refinement

Rule 3: A class with at least $x\%$ more manual annotations than its siblings is a candidate for refinement

with x taking four evenly spaced values between 100 and 250%.

Distinguishing between these two sets of annotations is very relevant in the context of GO, since the set of manual annotations contains only those that have been reviewed by a curator and can therefore be considered more

reliable. Nevertheless, only about 3% of all annotations are manual which means they provide a narrower coverage.

We applied these rules to classes across the 12 ontology versions. To accomplish this, we checked how well the two sets of classes created by the application of each rule, reflected the sets of classes that were refined and not refined in a future version at 6 months, 1 and 2 years. To evaluate the predictive power of the rules, we computed the number of true positives, true negatives, false positives and false negatives, and used the following indicators:

$$precision = \frac{tp}{tp+fp} \quad recall = \frac{tp}{tp+fn}$$

$$f - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Table 1 shows these results for refinement in 6 months¹ for the values of x that generated the best results.

Although these results are overall poor, there is a marked difference between the performance of structure and data-based rules, with data-based rules having a higher precision for all 3 hierarchies and a higher recall in molecular function.

We also applied these rules to predicting the refinement for ontology branches as a whole, as opposed to the previous strategy that predicted refinement for individual classes. This follows from the observation that many of the new classes inserted in the biological process hierarchy are inserted as part of small subgraphs rather than single insertions. We focused on the subgraphs that are rooted on classes at a depth of 4 due to the fact that most extension events occur at this depth or lower. However, the results obtained were comparable to those generated by predicting for individual classes.

4 Discussion

The application of our framework for ontology extension analysis to GO has yielded some interesting results. Firstly, the majority of new classes are not added to leaf classes, resulting

¹ Results for 1 and 2 years were similar; data not shown.

in a horizontal growth of the ontology. This means that GO is not adding increasingly specific classes but rather fleshing out. Secondly, we have identified that in GO refinement happens by two major modes: individual insertions and group insertions. The first occurs frequently in all GO hierarchies, whereas the second is only common in the biological process hierarchy. This is in line with the fact that most of GO's special interest groups belong to the biological process area and their work is more focused on modelling portions of their areas of interest rather than making individual insertions. We are aware that our usage of path-based depth to define the sub-graphs of GO that are subject to extension, can suffer from bias, since terms at the same depth do not necessarily express the same degree of specificity [1]. However, we have decided to use path-based depth, since we needed to create sub-graphs independently of their number of annotations, so as not to introduce a bias to our annotation based rules.

Biological Process			
Rule	Precision	Recall	F-measure
1 ($x = 75\%$)	0.0772 \pm 0.0317	0.364 \pm 0.0802	0.127 \pm 0.0479
2 ($x = 200\%$)	0.220 \pm 0.0185	0.318 \pm 0.0638	0.256 \pm 0.0128
3 ($x = 200\%$)	0.242 \pm 0.0302	0.380 \pm 0.0507	0.292 \pm 0.01714

Cellular Component			
Rule	Precision	Recall	F-measure
1 ($x = 75\%$)	0.0270 \pm 0.0228	0.381 \pm 0.206	0.0501 \pm 0.0406
2 ($x = 200\%$)	0.119 \pm 0.109	0.212 \pm 0.246	0.149 \pm 0.148
3 ($x = 200\%$)	0.199 \pm 0.121	0.374 \pm 0.259	0.252 \pm 0.156

Molecular Function			
Rule	Precision	Recall	F-measure
1 ($x = 75\%$)	0.0122 \pm 0.0033	0.223 \pm 0.0908	0.0230 \pm 0.0060
2 ($x = 200\%$)	0.101 \pm 0.0388	0.406 \pm 0.0357	0.157 \pm 0.0492
3 ($x = 200\%$)	0.123 \pm 0.0515	0.526 \pm 0.0573	0.194 \pm 0.0672

Table 1. Prediction results for the refinement of the Gene Ontology at 6 months. Shown values are averaged over all ontology versions, resulting from a total of 11 runs.

This refinement by branches in the biological process hierarchy is also captured by the enrichment analysis, where there is a high proportion of new enrichment relations that are established between new classes.

Theoretically, these two modes of refinement should impact semi-automated change capturing methods, hence we applied the rules for both individual and branch extension prediction. However such impact was not visible, likely due to the poor

performance obtained.

These results emphasize that the current proposed guidelines for capturing change based on structure and data are not appropriate for handling a large and complex ontology such as the Gene Ontology. We are aware that the guidelines represent an effort to ensure a balanced structure for the ontology, and that given the size and evolving nature of the domain GO covers, its extension cannot be governed alone by these precepts. In fact, GO's Ontology Development group² has highlighted the processes used in the identification of areas that need to be developed:

- by working closely with the reference genome annotation group to ensure that areas that are known to undergo intense annotation in the near future are updated
- by listening to the biological community
- by ensuring that emerging genomes have the necessary classes to support their needs

If GO's change management regarding extension were to be made explicit, for instance as is the case for making a term obsolete where the reason is given, we could perform a more in-depth analysis and perhaps derive more accurate rules. Nevertheless we have obtained better results using the number of annotations rather than the number of subclasses, which may be related to the fact that GO development is driven by need, which can be approximated by the rate of annotation, rather than by a process of homogenization of structure. In fact, this difference was to be expected considering that in GO's domain the level of specificity of each branch is dependent on natural and scientific phenomena, which prevents the existence of a homogenous structure to the ontology. Such structure-based guidelines are however expected to function better in ontologies that follow a more conceptual approach.

In trying to predict ontology extension, particularly in the case of large biomedical ontologies, we are facing a multitude of variables, not only the advancement of biomedical knowledge and the current state of the

² http://wiki.geneontology.org/index.php/Ontology_Development_group_summary

ontology itself but also social and technical aspects. The extension of biomedical ontologies occurs via several different processes, and motivated by distinct needs, which cannot be apprehended by a 'one size fits all' rule. We believe that to handle this complexity, we need to employ more sophisticated techniques that are able to handle numerous variables and more complex relations between them.

Therefore we are currently working on a supervised learning methodology to support the prediction of ontology extension that explicitly addresses these issues.

We believe that the future of ontology development will necessarily incorporate the automation of some of its processes, mainly those that are tedious and time-consuming, releasing ontology experts to focus on core modelling issues. We have outlined one of these processes, semi-automated change capturing via prediction of ontology extension and presented some of the issues and challenges in this budding field.

Acknowledgements

This work was supported by FCT, through the Multiannual Funding Programme and the PhD grant SFRH/BD/42481/2007. The authors also wish to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807).

References

1. G. Alterovitz, M. Xiang, D. P. Hill, J. Lomax, J. Liu, M. Cherkassky, J. Dreyfuss, C. Mungall, M. A. Harris, M. E. Dolan, J. A. Blake, and M. F. Ramoni. Ontology engineering. *Nature biotechnology*, 28(2):2008–2011, 2010.
2. S. Castano, A. Ferrara, and G. Hess. Discovery-driven ontology evolution. *The Semantic Web Applications*, 2006.
3. W. Ceusters. Applying evolutionary terminology auditing to the Gene Ontology. *Journal of biomedical informatics*, 42(3):518–29, 2009.
4. W. Ceusters and B. Smith. A realism-based approach to the evolution of biomedical ontologies. *AMIA Annual Symposium Proceedings*, (4):121–5, Jan. 2006.
5. P. Cimiano and J. Völker. A framework for ontology learning and data-driven change discovery. *Proc. of the NLDB2005*, 2005.
6. G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, and G. Antoniou. Ontology change: classification and survey. *The Knowledge Engineering Review*, 23(02):117–152, 2008.
7. J.-B. Lee, J.-J. Kim, and J. C. Park. Automatic extension of Gene Ontology with flexible identification of candidate terms. *Bioinformatics*, 22(6):665–70, 2006.
8. P. D. Leenheer and T. Mens. Ontology evolution: State of the Art and Future Directions. *Ontology Management*, 2(1):1–47, 2008.
9. V. Nováček, L. Laera, S. Handschuh, and B. Davis. Infrastructure for dynamic knowledge integration- automated biomedical ontology extension using textual resources. *Journal of biomedical informatics*, 41(5):816–28, Oct. 2008.
10. N. F. Noy and D. L. McGuinness. Ontology Development 101 : A Guide to Creating Your First Ontology. *Development*, pages 1–25, 2000.
11. F. Silva, M. Silva, and F. Couto. Epidemic Marketplace: an e-Science Platform for Epidemic Modelling and Analysis. *ERCIM News*, 2010.
12. L. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic. User-driven Ontology Evolution Management. *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW-02), Lecture Notes in Computer Science (LNCS), Volume 2473, Springer-Verlag*, pp:285–300, 2002.
13. L. Stojanovic and B. Motik. Ontology Evolution Within Ontology Editors. *Proceedings of the OntoWeb-SIG3 Workshop*, pp:53–62, 2002.
14. T. Wächter and M. Schroeder. Semi-automated ontology generation within OBO-Edit. *Bioinformatics*, 26(12):i88–96, June 2010.

Revising the Cell Ontology

Terrence F Meehan¹, Christopher J Mungall², Alexander D Diehl³

¹Mouse Genome Informatics, The Jackson Laboratory, USA; ²Lawrence Berkeley National Laboratory, USA;

³Department of Neurology, University at Buffalo School of Medicine and Biomedical Sciences, USA

Abstract. The Cell Ontology (CL) is an ontology of in vivo cell types that is undergoing extensive revision to become a full member of the OBO Foundry. To help us achieve this, a series of goals was established at a CL development workshop in May 2010. Here we describe our ongoing efforts to meet these goals including the modification of the CL's domain, the import of over 400 cell types from the Foundational Model of Anatomy, the addition of both free text and logical definitions, and the incorporation of new terms in response to our user community. These enhancements increase the utility of the CL for both researchers and ontology developers while adhering to the principles of the OBO Foundry.

Keywords: Cell Ontology, CL, OBO Foundry

1 Introduction

The Cell Ontology (CL) is a candidate OBO Foundry ontology for the representation of cell types. First described in 2005 [1], the CL from its earliest incarnation has attempted to integrate cell types from the prokaryotic, fungal, animal and plant organisms. The original developers felt the advantages of having a common framework outweighed the difficulties in incorporating cell types from different phyla. As a core component of the OBO Foundry, the CL merges information contained in species-specific anatomical ontologies as well as referencing other OBO Foundry ontologies such as the Protein Ontology (PR) for uniquely expressed biomarkers and the Gene Ontology (GO) for the biological processes a cell type participates in [2,3].

An area of the CL that has benefited from the ongoing development of another ontology is the hematopoietic cell branch. In conjunction with a revision of immunological processes in the GO, about 80 new immune cell types were added to the CL in 2006 [4]. This brought the CL to the attention of the National Institute of Allergy and Infectious Disease, which sponsored a workshop in 2008 that brought together domain experts and biomedical ontologists to further improve immune cell representation in the CL [5].

Besides identifying missing cell types, participants agreed that creating logical definitions for cell types built from relationships to other OBO ontologies would increase accuracy and interoperability with other ontologies. Based on the experts' input, logical definitions (also known as computable definitions or cross-product definitions) were first created for dendritic cell types [6] and then subsequently for all hematopoietic cell types [7]. This approach has not only increased accuracy of the ontology but has also led to unexpected associations between cell types suggesting the ontology could be used for hypothesis generation. Building on this success, we have begun revising the whole of CL. To help with this, a workshop was convened with ontology experts in May 2010 and several goals were set for the further development of the CL. Here we describe our efforts to meet these goals and revise the CL for entry into the OBO Foundry.

2 Method

2.1 Cell Ontology Development Workshop

A workshop was held at The Jackson Laboratory on May 18-19, 2010. Participants included OBO Foundry ontology developers and users of the CL. The goal of this workshop was how to best resolve the problems of the

CL in relation to the OBO Foundry Principles [8] while meeting the requirements of our funding. A summary of the workshop can be found here: http://obofoundry.org/wiki/index.php/Cell_Ontology_Workshop_2010.

2.2 Editing and Generating Different Versions of the CL

The Cell Ontology has been developed as an OBO format ontology using OBO-Edit software [9]. CL ontology developers modify an editors' version of the ontology in the file, `cell.edit.obo`, which contains the CL as an unreasoned ontology that includes the minimum and necessary classes (i.e. MIREOTED classes [10]) from other OBO ontologies to allow for sufficient reasoning. An OWL version of the CL, `cell.edit.owl`, is generated from `cell.edit.obo` using the standard `obolib-obo2owl` converter (<http://code.google.com/p/oboformat>). This converter also performs macro-expansion of shortcut relations, as previously described [7] (also in <http://www.berkeleybop.org/~cjm/obo2owl/obo-syntax.html>).

An example of a shortcut relation is

lacks_plasma_membrane_part

which is used to define a cell type by an absence of cell surface marker, and has the following macro definition:

has_part exactly 0 (GO:'plasma membrane' and has_part some ?Y)

CL editors used the fully expanded ontology to find errors and make corrections to the `cell.edit.obo` file. A pre-reasoned version of the CL, `cell.obo`, is generated from `cell.edit.obo` and has all implied links asserted and MIREOTED classes removed for full compatibility with existing tools.

CL is available in two separate forms in

either of two formats from the URLs below.

2.3 Obol Analysis and Import of Foundational Model Anatomy Classes

The FMA contains a number of cell type classes that are locationally qualified, for example “mesothelial cell of visceral pleura” and “endothelial cell of hepatic sinusoid”. To support our goal of making the Cell Ontology the central repository of cell types, we set about generalizing these classes and placing them in CL. We used the Obol tool [11] to parse the labels of the FMA classes into logical definitions, typically consisting of a generic cell type and a gross anatomical location qualified by the `part_of` relation (for example, “mesothelial cell” and `part_of` some “visceral pleura”). We mapped the generic cell type to a CL class, and the location to an Uberon class. The mappings were done on the basis of existing `dbxrefs` maintained in CL or Uberon. Where no generic cell class existed in CL, we manually created one in OBO-Edit.

3 Results

3.1 Providing Different Versions of the CL

The first priority set by members of the CL workshop was to extend our use of logical definitions for hematopoietic cell types to the whole of CL. A logical definition is constructed in a modular fashion by using relationships to classes from other ontologies. These computable definitions can be expressed in ontology formats and languages such as OBO or OWL, and are treated as equivalence relationships between the defined class and some conjunction of classes. For example, the class “pancreatic centro-acinar cell” can be defined as equivalent to the class of things that both are “epithelial cells” and are part of the “pancreatic acinus”.

	OBO Format	OWL RDF/XML
Standard	http://purl.obolibrary/obo/cl.obo	http://purl.obolibrary/obo/cl.owl
Basic	http://purl.obolibrary/obo/cl-basic.obo	http://purl.obolibrary/obo/cl-basic.owl

Table 1: Download options for CL

All forms are pre-reasoned. The standard form includes MIREOTed classes, which could cause problems for some tools, we we also provide a basic form that has all MIREOTed classes and references to these classes removed.

We have recently published our work on generating logical definitions for the vast majority of hematopoietic cell types [7] and are continuing to use logical definitions as we add classes. Currently 586 of 1559 CL classes have a logical definition (Table 1). Of these definitions, 442 use macro relations that expand to more complex expressions that can be used by OWL reasoners [7] and are available in the OWL serialization of the ontology. We found that the use of macro relationships was critical in maintaining the logical structure of the CL because some inferences are incomplete with the rule based reasoner (RBR) in OBO-Edit. For example, the cell class “erythroid lineage cell (CL:0000764)” was originally defined with:

“myeloid cell *has_plasma_membrane_part*
transferrin receptor protein 1
(PR:000001945)”

while one of its descendent classes “erythrocyte (CL:0000232)” was partially defined as:

“erythroid lineage cell *lacks plasma membrane part* transferrin receptor protein 1”

The reasoners in OBO-Edit are unable to detect this logical inconsistency. However, with an OWL translation of the ontology and the use of macro relations (see Methods), OWL reasoners in Protégé identify the cardinality violation. We then changed the ontology accordingly.

We adjusted our workflow to take advantage of these macro relationships (Figure 1). For a typical user, we felt a pre-reasoned ontology that contained all links inferred by an ontological reasoner as fully asserted links and did not contain classes from other ontologies was important for ease of use. This version is available as <http://purl.obolibrary.org/obo/cl-basic.obo> and is identical to the “cell.obo” file deposited in the obo library CVS repository, which is maintained for historic reasons. The basic version is also available as OWL. We also make available a complete pre-reasoned version

<http://purl.obolibrary.org/obo/cl.{obo,owl}>.

This version includes the full set of MIREOTed classes and equivalence axioms linking to these classes. The CL editors edit the cell.edit.obo file, which is not pre-reasoned. This file currently resides in the sourceforge CVS repository and is in the process of moving to googlecode. Other versions of the Cell Ontology including an OWL version are provided as described in the methods.

3.2 Incorporating Foundational Model Anatomy Cell Types into the CL

A second important goal for the development of the CL was a reiteration that CL is the ontology for all *in vivo* cell types in the OBO Foundry. Thus, a cell type that is represented in an anatomical OBO ontology should have a species-neutral equivalent in the CL and a mapping between classes. As such, we modified the ontology so that a CL class is a superclass of all the classes stated as dbxrefs. For example, the CL class “photoreceptor cell (CL:0000210)” that describes any animal cell that is able to detect light and transduce a signal, has dbxrefs to equivalent classes in both fly anatomy and human anatomy ontologies. The dbxrefs are translated to *is_a* (SubClassOf) relationships between the CL and the respective classes in the different ontologies. Existing dbxrefs in CL that did not fit this criterion were moved to the comments section for the cell type. Once this was done, we examined the other OBO anatomical ontologies and decided to work first with the Foundational Model Anatomy ontology (FMA) for several reasons. First, the number of cell types represented in the ontology was in the hundreds compared to the thousands represented in some other ontologies. Second, the cell types in FMA contained many general cell types like “endo-epithelial cell” that were not present in the CL. Third, the FMA is a mature ontology that has been in development for over ten years. While the FMA is still being actively developed, major revisions in the ontology were not planned during the months it required us to import the cell classes.

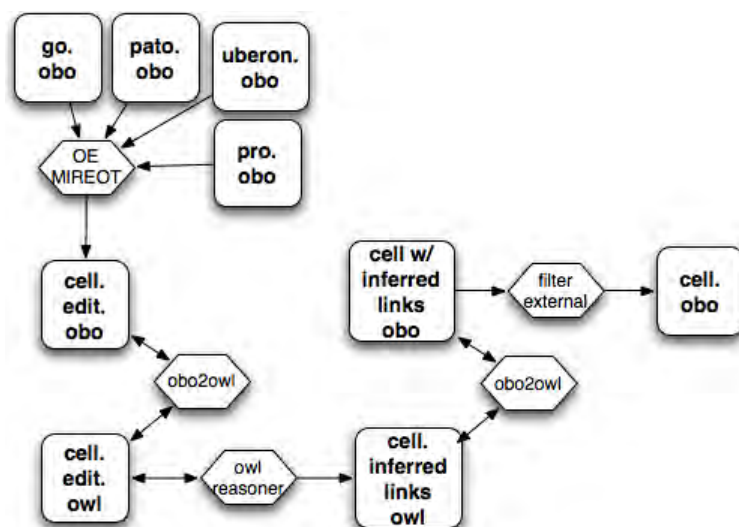


Figure 1. Dataflow for CL. External ontologies are incorporated using the oboedit MIREOT implementation. The obolib-obo2owl tool is used to interchange back and forth between obo and owl. Owl reasoners are used to generate an OWL file with inferred links materialized. The file is converted back to obo, where the file is filtered for the basic version.

We employed a term structure parsing tool called Obol to identify cell types in the FMA [11]. This approach gives unique non-trivial definitions to classes based on implicit rules inherit in the class names. The Obol analysis was able to generate logical definitions for 221 FMA cell types based on the syntax of: “<generic> of <anatomical entity>”. In 189 cases, a successful match was made from the “generic” FMA cell type to a pre-existing CL class. The missing 32 generic FMA cell types were then manually added to CL. This information was also supplied to the developers of the Uberon ontology [12]. Uberon is a multi-species anatomy ontology created to facilitate comparison of phenotypes across multiple species. As such, the Uberon developers referenced the species-neutral anatomical structures in their ontology to the corresponding structures in the FMA. Once these mappings were complete, the 221 cell types parsed by the Obol analysis were added automatically to the CL complete with logical definitions using Uberon classes and dbxref to the FMA class.

Obol analysis of the FMA was unable to parse 539 classes beyond the classification of “cell type”. Of these, 213 classes had dbxrefs references in the CL and 88 were neuronal cell types, which were put aside for a future workshop (see “Discussion”). The remaining 238 classes were reviewed in detail and ultimately over 200 new cell types were added

with free text definitions to the CL. Cell types ranged from those type that failed to parse in the Obol analysis due to historical names like “Boettcher cell” to cell types whose names reflected their highly specialized nature like “type II cell of carotid body”. While laborious, addition of these cell types added great value to the CL including many cell types performing unique biological roles that will aid in logical definitions of the Gene Ontology. An important addition to the CL was cell types that reflect lineage development such as “endo-epithelial” defined as “An epithelial cell derived from endoderm.” 83 cell types have this term as an ancestor, which means all these cell types can be traced back to the endoderm lineage (Figure 2). This greatly extends the representation of developmental lineages in the CL and will serve as a cornerstone as we further develop this aspect of the ontology. By placing FMA cell types in the context of the CL and using logical definitions, we could find equivalent cell types. For example, Type F (FMA:83409) and Type PP (FMA:62938) enteroendocrine cells were found to be equivalent as both cell types are defined by secreting pancreatic polypeptides.

Developers of the FMA are considering obsoleting the cell-type classes in favor of the CL classes. Until this is done, bridging between FMA and CL will be provided by our dbxrefs where a reference to a FMA class represents the human equivalent of cell-type represented by a species neutral CL class.

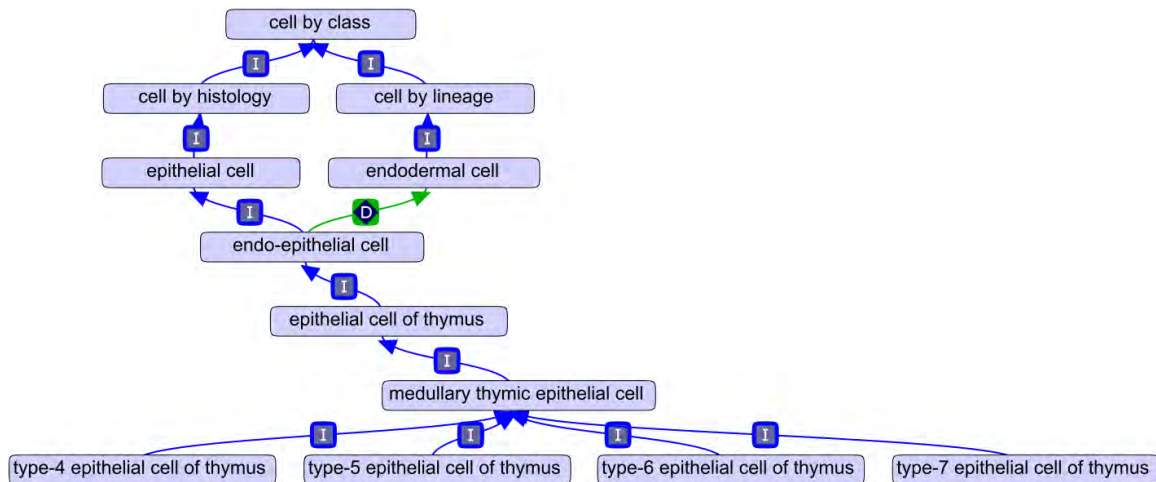


Figure 2. Endo-epithelial cell type and a subset of its descendents.

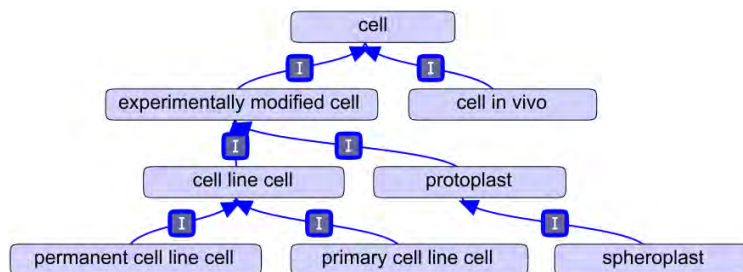


Figure 3. Descendants of experimentally modified cell will be removed from CL

3.3 Removal of Experimentally Modified Cell Types

Another goal established at the Cell Ontology workshop is that experimentally modified cell types should be moved into OBI. When the CL was implemented, *in vitro* cell types were included in its domain such as “primary cell line” and “permanent cell line” (Figure 3). Since then, the Ontology of Biomedical Investigations (OBI) has been undergoing active development and many of the workshop participants felt *ex vivo* cell types and cell lines fall under OBI’s domain. OBI is a candidate OBO Foundry ontology that represents the design, materials, implementation, and data of biological investigations [13]. As cell lines are experimentally derived entities, they arguably fall within OBI’s domain. Biologists agree for multiple reasons that cell lines often bear little resemblance to the cell types they are derived from. For example, principal component analysis of a large number of gene

expression arrays demonstrates that cell lines fail to cluster with their tissues of origin and instead tend to cluster among themselves [14]. While there was general agreement that cell line classes be moved to OBI, many participants felt that mappings between cell lines and the CL should be made. Towards this goal, we have been working with the developers of a cell line ontology that is based on the Cell Line Knowledgebase [15]. We have begun adding missing cell types to the CL from which these cell lines are derived. These additions include logical definitions to Uberon classes to provide a generalized anatomical context. A *derives_from* relationship is used in the cell line ontology to the appropriate CL class. More about this work appears in a related ICBO submission.

3.3 Other Developmental Goals for the CL

Several other goals were established for the development of the CL. One goal is to provide free text definitions and biomedical references for all CL classes. Currently 1263 classes of

1559 have some sort of free text definition with the majority of those containing references to the biomedical literature (see Table 1). Another goal was to improve response time for requested changes in the CL. Excluding neuronal classes (see “Discussion”), we have worked through a three-year backlog of tracker items on the CL SourceForge tracker and now respond to most term requests within a week. Other accomplishments achieved since the workshop include the import of our sub-ontology of hematopoietic cells (called Hemo_CL) [7] into the CL, providing better documentation of the CL structure and development, and increasing our outreach to potential users of the CL ontology.

4 Discussion

Here we describe our improvements in the representation of cell types in the Cell Ontology that enhances its use in data annotation and integration. By implementing these changes, the CL is taking on a core role in the OBO Library by serving as a conduit to link anatomical ontologies to the GO, the PR and other OBO ontologies. We have done this in a manner that adheres to the principles of the OBO Foundry, namely by working collaboratively with other OBO Foundry members to provide a clearly delineated ontology expressed in both OBO and OWL formats that is available for use by all.

Ontology	cell.edit.obo v1.1	cell.edit.obo v1.62
Cell type classes	988	1559
Classes with free text definitions	565	1263
Classes with logical definitions	0	586
Relationships used with CL classes	2	10
Number of External OBO classes MIREOTED into the CL	0	1005
Number of xrefs	121	564

Table 2. Summary of changes in the CL from September, 2009 through January, 2011

Perhaps the most important goal for development of the CL is to continue outreach to both ontology developers and to end-users. Much of the work described herein stems from CL development workshops that involved biomedical researchers. Thus, the CL has become more reflective of the needs of our users. For example, we have recently been asked to join the FANTOM5 consortium, which seeks to map transcription initiation in over 200 human cell types. The organizers felt the CL’s representation was broad and deep enough to help structure the terabytes of data the project is expected to generate, and we have committed to adding additional cell types needed for the FANTOM5 work. Outreach also helps coordinate the CL with other OBO Foundry ontologies. We have recently held a workshop to extend the representation of neuronal cell types in collaboration with the International Neuroinformatics Coordinating Facility (INCF). The workshop included experimental neuroscientists including members of the INCF Neuron Registry Task Force and developers of anatomical ontologies for Human, Drosophila, rodent, and the pan-species Uberon ontology. Despite differences in neuronal representation by these ontologies, we believe a unified approach to neurons is possible based on our past experience with having ontology developers and biologists work together and we have now begun curating neurons based on this approach. This in turn will help integrate data in the respective ontologies.

In summary, we continue to develop the CL to enhance its usefulness for researchers and ontology developers, and for consideration as a full member of the OBO Foundry.

Acknowledgments

The Cell Ontology project is supported by an NHGRI-funded, ARRA administrative supplement grant HG002273-09Z to the parent grant, HG002273, to the Gene Ontology Consortium. This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Bard J, Rhee SY, Ashburner M. An ontology for cell types. *Genome Biol.* 2005;6(2):R21.
2. Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, et al. Cross-product extensions of the Gene Ontology. *J Biomed Inform.* 2011 Feb;44(1):80-86.
3. Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, et al. The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res.* 2011 Jan;39(Database issue):D539-545.
4. Diehl AD, Lee JA, Scheuermann RH, Blake JA. Ontology development for biological systems: immunology. *Bioinformatics.* 2007 Apr 1;23(7):913-915.
5. Diehl AD, Augustine AD, Blake JA, Cowell LG, Gold ES, Gondré-Lewis TA, et al. Hematopoietic cell types: prototype for a revised cell ontology. *J Biomed Inform.* 2011 Feb;44(1):75-79.
6. Masci AM, Arighi CN, Diehl AD, Lieberman AE, Mungall C, Scheuermann RH, et al. An improved ontological representation of dendritic cells as a paradigm for all cell types. *BMC Bioinformatics.* 2009;10:70.
7. Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, Mungall CJ, et al. Logical development of the cell ontology. *BMC Bioinformatics.* 2011;12:6.
8. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 2007 Nov;25(11):1251-1255.
9. Day-Richter J, Harris MA, Haendel M, Lewis S. OBO-Edit--an ontology editor for biologists. *Bioinformatics.* 2007 Aug 15;23(16):2198-2200.
10. Courtot M, Gibson F, Lister AL, Malone J, Schober D, Brinkman RR, et al. MIREOT: The minimum information to reference an external ontology term. *Applied Ontology.* 2011 Jan 1;6(1):23-33.
11. Mungall CJ. Obol: integrating language and meaning in bio-ontologies. *Comp. Funct. Genomics.* 2004;5(6-7):509-520.
12. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* 2009 Nov;7(11):e1000247.
13. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, et al. Modeling biomedical experimental processes with OBI. *J Biomed Semantics.* 2010;1 Suppl 1:S7.
14. Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.* 2010 Dec 23;11(12):R124.
15. Sarntivijai S, Ade AS, Athey BD, States DJ. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics.* 2008 Dec 1;24(23):2760-2766.

Rapid Development of an Ontology of Coriell Cell Lines

Chao Pang, Tomasz Adamusiak, Helen Parkinson, James Malone

European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK

Abstract. Many online catalogues of biomedical products and artifacts exist that are loosely structured but of great value to the community. These include cell lines, enzymes, antibodies, reagents, and laboratory equipment. Improving the representation of these products has several benefits; detailed querying power, product comparison and reporting of products used in experimental protocols. However, this formalization is often time-consuming, labor-intensive and expensive. We describe an approach to structuring these catalogues using semi-automated techniques to rapidly develop OWL ontologies. We demonstrate the approach using the Coriell Cell Line catalogue, and the resulting ontology of 28,000 classes which imports classes from other community ontologies such as Disease Ontology, Cell Type ontology and FMA. **Availability:** <http://efo.sourceforge.net/coriell.htm>

Keywords: Coriell ontology, automated ontology engineering, cell line ontology

1 Introduction

The biomedical community has embraced the use of ontologies as a means of describing scientific data, such as experimental protocols (OBI) [1] experimental variables (EFO) [2] and phenotypes [3]. The development of these ontologies, however, is a costly activity. It requires considerable time and expertise to produce a large and/or complex ontology.

There is clearly value in producing robust expertly curated ontologies. The Gene Ontology (GO) [4] is the archetypal example of this, is developed by a team of experts and is continuously updated to include new biological knowledge. However, development in this form is clearly not repeatable across every area of biomedicine. There is evidence that methods and tools that expedite the process of ontology engineering are much needed [5].

Programmatic approaches can be powerful when transforming resources with some pre-existing structure into an ontological form [6]. Loosely structured data sources contain implicit knowledge – within the data or within the presentation layer, for example within categories in a drop-down list on a website. This is often the case for data stored in a database which is accessible through a web interface, which may contain some logic behind the sorting of options but which is otherwise unavailable. Similarly, such implicit knowledge

may also be contained within the column headers of spreadsheets or within database table and field names. In such cases it may be possible to exploit implicit knowledge and develop models which enable a rapid transform into ontology classes.

In this paper we present our approach to the rapid development of the large Coriell cell line ontology based on a collection of semi-structured cell line descriptions from the Coriell cell line catalogue. The Coriell cell line catalogue contains ~27,000 mammalian cell lines and we demonstrate that by using a standardized modeling pattern and text mining approaches, a large ontology containing >28,000 classes can be rapidly produced which logically describes each cell line and their biological properties.

2 Methods

The principle methodology underlying this work is ontology normalization [7]. Specifically, that we manage multiple inheritance using class descriptions in OWL and infer structure using description logic reasoners such as HermiT [8]. By providing axioms on classes, the need to assert potentially conflicting or fragile subsumption hierarchies is removed. This approach also makes the biological knowledge used to create the hierarchy explicit and therefore renders implicit knowledge explicit in the ontology.

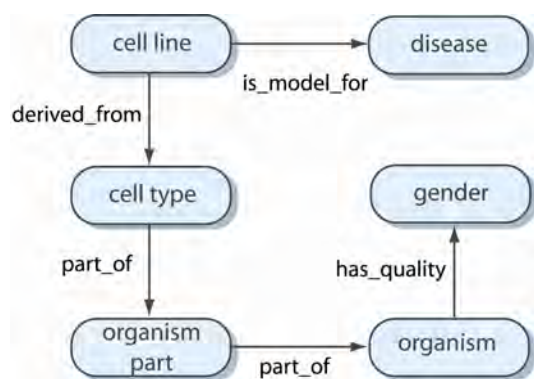


Figure 1. Part of the cell line model used to represent Coriell cell lines.

The first step in producing the ontology was to develop a standardized model for cell lines. By collaboration with the Cell Line Ontology [9] and the Cell Type Ontology [10] we created a model (Figure 1) which aligns these ontologies and which was used during development of the Coriell Cell Line Ontology.

Our primary queries of interest are contained in this model, specifically: cell line name, cell type, disease, organism parts, organism and gender. The model therefore represents the key attributes of Coriell cell Lines and was evaluated against primary competency questions derived from use cases related to the development of a BioSample Database (www.ebi.ac.uk/biosamples/) at the EBI. These include queries by common cell types, by disease and tissues. We use a 'short cut' relation, *is_model_for*, to associate cell lines with disease, this reflects the use of cell lines as models for particular diseases. Given the large size of the Coriell catalogue we developed a scalable semi-automatic approach to creating the ontology. Information on each cell line was contained within 104 separate and redundant text files each describing different aspects of the Coriell products and derived from an SQL dump of a relational database. Five key files were selected which contained semi-structured descriptions covering the entities described in Figure 1 and which corresponded to our use cases. These files were merged, redundant information was removed and a single 'cell line' spreadsheet was produced using bespoke Perl scripts.

2.1 Lexical Entity Mapping

The cell line spreadsheet was used as an input

for lexical entity mapping with the aim of generating list of classes from reference ontologies that matched the textual descriptions. The Perl OntoMapper [2] was employed as it has previously been used successfully in building similar application ontologies. The approach allows for fuzzy matching to identify classes from class labels and their synonyms. Given the nomenclature of areas such as disease and anatomy where synonymy is common, a fuzzy matching approach provided flexibility in mapping. A metric was assigned to each match and those with less than 100% confidence were manually inspected.

The reference ontologies (Table 1) were selected based on the content of the files and the model. Anatomy was particularly challenging as although the Coriell cell lines were primarily mammalian no single mammalian anatomy ontology exists which would provide the coverage necessary. Although some efforts are ongoing to develop an homology based anatomy ontology [11, 12] we used a pre-existing resource the Minimal Anatomy Terminology [13]. This species neutral ontology provides mappings to multiple anatomical ontologies and is subsumed by the Experimental Factor Ontology, with which we plan to merge the Coriell Cell Line Ontology in future. When a core mammalian anatomy ontology becomes available we will replace the MAT. Some human specific classes were also imported from FMA.

The disease information within the Coriell descriptions consisted of references to OMIM [14]. Since OMIM is not a disease ontology we exploited the links provided within the Human Disease Ontology (DO) to OMIM and imported DO classes. Where links were not made between OMIM and DO, a manual inspection using BioPortal [15] was required to extract the corresponding disease.

Domain	Reference Ontology	Term Number
Organism	NCBI Taxonomy, OBI	93
Anatomy	Experimental Factor Ontology, FMA	61
Cell Type	Cell Type Ontology	11
Disease	Human Disease, NCI Thesaurus	337
Gender	PATO	3

Table 1. Reference ontologies used in the Coriell cell line ontology.

2.2 Ontology Engineering Using OWL-API

The lexical mapping resulted in a set of files containing mappings between a label and the corresponding URI from the reference ontology, one file per domain. These mappings were used to construct the ontology programmatically – Figure 2 illustrates this process.

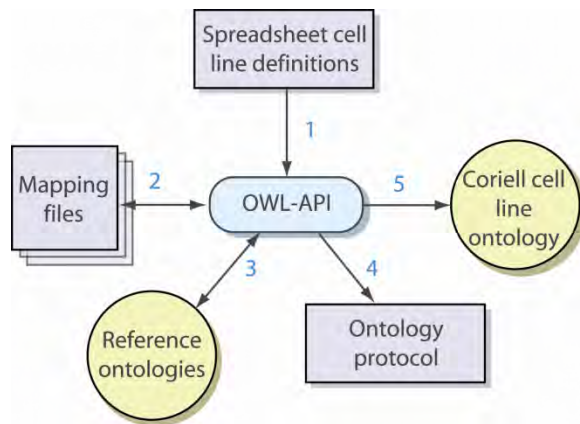


Figure 2. Methodology used for programmatic ontology creation

The process was implemented as follows:

- (1) Input of cell line descriptions contained in the single merged spreadsheet.
- (2) Files containing mappings from class label to reference ontology class IRI (Internationalized Resource Identifier) are matched.
- (3) Class IRIs are used to import corresponding ontology classes from reference ontologies, along with axiomatic and annotation information within the class signature if present and parent classes.
- (4) The EFO upper level is re-used here (a slim version of BFO) and determines where imported classes should be placed. For example, disease classes are imported under the *disease* parent, itself a child of *disposition*. This protocol also determines how axioms on classes should be formed.
- (5) The Coriell cell line ontology in OWL is output.
- (6) The ontology was manually reviewed for correctness, checked for consistency using Hermit 1.3.1 and defined classes were added to ensure axioms were used,

formulated correctly and meaningful as well as to add structure to the hierarchy.

3 Results

The Coriell cell line ontology produced contains 27,002 cell line classes, covering 11 cell types, 61 anatomical terms and 93 organisms. 657 OMIM numbers were attached to cell lines and 393 OMIM numbers were mapped to 337 unique Disease Ontology classes. 7,688 cell lines were confirmed to model disease and a small number modeled multiple diseases, for example ND00139 which models Parkinson's Disease and Lewy Body Disease.

Following the creation of the ontology some refinements to the imported structure were required.

3.1 Organism Taxonomy

Organism classes imported from the NCBI taxonomy [16] have very long chains of parent classes. For example *Homo sapiens* has 28 classes in a subclass hierarchy between it and the parent class organism. We then retrospectively removed some of these nodes, applying the following design principle; 1. Remove intermediate classes when the child class does not have more than 2 siblings, 2. When the deletion leads to >3 child classes, the parent class is retained. This strategy removed a large number of classes which were not required by our query use cases and these could easily be added back in if needed in future.

3.2 Anatomy

There were 81 unique terms describing anatomy, 45 mapped exactly to pre-existing terms in the MAT. Unmapped terms describe classes other than anatomy such as fibroma, leiomyoma (which are disease classes) and were removed. Buttock-thigh and Thorax/abdomen could be separated into two single terms but it is not clear which part the terms were describing and these were also removed. 9 terms were unmapped which did not appear to fit into anatomy, such as Keloid breast organoid, so were removed. Among the remaining terms unmapped from the entity recognition step, 12 terms are mapped to FMA, 9 terms to EFO, 2 terms to SNOMED CT ontology, 2 terms to

NCI Thesaurus and one term is not mapped to any ontology. The mixing of terms from disease and anatomy domains was found to be common in many parts of the Coriell Catalogue and manual effort was spent assessing these terms prior to building the ontology.

3.3 Cell Type

22 unique cell type terms were mapped to the Cell Type Ontology. 11 terms are with 100% similarity. Partial mappings were refined manually e.g. *smooth muscle* is not a cell type and was modified to *smooth muscle cell*. Myeloma is not a cell type, but a cancer of plasma cells and was changed to plasma cell. Another 11 unmapped terms were not cell type terms and were removed.

3.4 Disease

We use the structure described in Disease Ontology in the Coriell cell line ontology and imported 337 disease terms. DO, it is not axiomatised except for the use of subclass relationships. EFO, however, provides more information for the class relationships (e.g. disease to anatomical parts). For disease we therefore added axioms to allow construction of defined classes based on e.g. disease e.g. '*liver disease cell lines*'. Firstly DO classes mapped from the cell line description spreadsheet were imported including any DO metadata. Then imported classes were axiomatised using additional logical restrictions from EFO (e.g. an axiom linking disease to anatomical part). The axiomatic information imported from EFO does not affect the DO child and parent classes and therefore the canonical structure from DO (and DO IRIs) is preserved.

3.5 Adding Defined Classes to Infer Structure

Use of normalisation methodology results in an asserted flat cell line hierarchy, i.e. the only asserted parent class of each cell line is the cell line class. For browsing purposes, however, it is often useful to produce an organizational hierarchy and as such we created some under cell line using defined classes in OWL, i.e. classes with necessary and sufficient restrictions describing members.

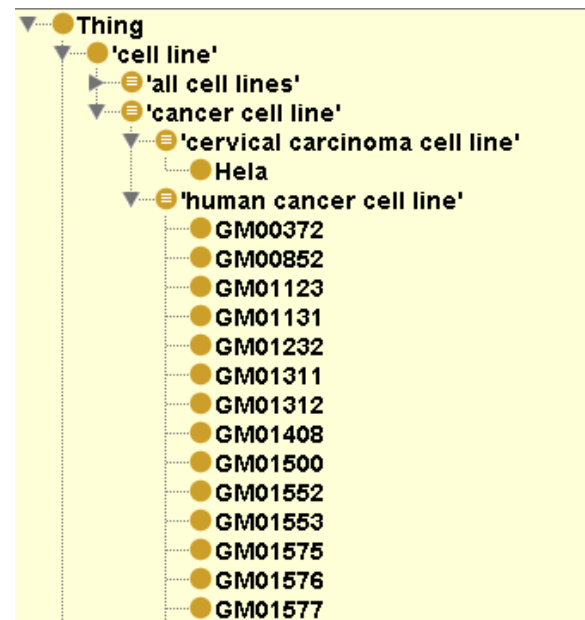


Figure 3. Inference of cell line hierarchy shown in Protégé.

The defined class *human cancer cell line* has necessary and sufficient conditions which result in various cell lines inferring as subclasses following the use of a description logic reasoner. This is very useful in rapidly creating dynamic hierarchies which can be changed very easily and for managing multiple inheritance.

For example, human cancer cell line (Figure 3) has the following necessary and sufficient restriction using Manchester OWL syntax [17]:

```

'cell line'
and (is_model_for some cancer)
and (derives_from some
('cell type'
and (part_of some
('organism part'
and (part_of some 'Homo sapiens')))))

```

The nesting reflects an important distinction between separate statements; in effect, we are saying for a specific *organism*, for which a specific *organism part* is part, and from which a specific *cell type* was taken. For the example in Figure 3, the defined class restricts membership to those classes where cancer is the modeled disease and which are derived from humans (more specifically cell types that are part of an organism part which are part of humans).

We have also used disjoints in some areas of the ontology, for example by making Homo

sapiens disjoint from other siblings under organism, we are able to ask the query for things which are not Homo sapiens because they have been explicitly stated as such. The following returns a class of cell lines that are not derived from human:

```
'cell line'
and (derives_from some
('cell type'
  and (part_of some
    ('organism part'
      and (part_of some
        (organism
          and (not ('Homo sapiens'))))))))
```

3.6 Rapid Generation and Rapid Regeneration

The ontology was developed over 3 months by one person working full time. The majority of this time was spent developing the code to produce the ontology and a repeat exercise using similar methods would take a great deal less. We made several changes to the ontology as we progressed and refined the model slightly; the programmatic method used meant regenerating the new OWL ontology took minutes. Moreover, rapidly adding new content programmatically is also possible.

4 Discussion

One of the central claims of this work is that the ontology was rapidly developed using the methods described. Over the 3 months that this work was conducted, we estimate 2 months comprised investigation of the catalogue content and Perl scripting to merge and format the initial input files. A further month's programming resulted in an ontology of ~28,000 classes. Generalizable components of the methodology include: design of reusable design patterns, re-use of ontology development code and exploitation of the MIREOT process for term imports.

There is a trade-off between hand crafted curation by individual experts and the rapid development of a very large resource. Our approach is of most benefit when a semi-structured data exists and existing Foundry type ontologies are available e.g. for cell types. As a one-off SQL dump was used for development updates need to be managed in

future and a dynamic method for accessing new data is desirable.

One of the criteria for inclusion in the OBO Foundry effort [18] is that every class is given a textual definition. The effort required to manually produce good textual definitions for an ontology the size of the Coriell cell line ontology is significant. Given the axiomatisation of the ontology, however, efforts such as producing natural language from OWL statements may offer an effective and rapid method to producing textual definitions [19]. If such an approach can be applied we will seek to include the artifact into the OBO Foundry in the future. We are also currently working with the Cell Line Ontology to ensure our respective models are synchronized and to merge the Coriell cell line ontology with the CLO which is currently derived from the American Tissue Culture Collection (ATCC). Other work includes mapping to all resources which contain cell line references and addition of these to the ontology, re-running of imports to detect changes in source ontologies, term requests from e.g. the cell type ontology to classify cells by anatomical part and addition of information manually where possible e.g. much text containing phenotypic descriptions was unstructured and could be mined added. A complete evaluation of additional meta data vs. that of the CLO is also desirable in order to prioritise where to add curation effort and which additional data could added to the core we have built. This work has allowed us to refine the cell line model within EFO to be consistent with the CLO and this will be revised in future releases of EFO. Future work also includes the release of the Coriell ontology to Bio2RDF for linked open data access. Finally our programmatic approach is fully compatible with manual curation and ontology development, and a combined approach is likely to produce rich, well structured ontologies for community use.

Acknowledgments

We thank the Functional Genomics Production Team, the Coriell Institute for Medical Research, Alan Ruttenberg and Science Commons for providing the Coriell SQL dump. Lynn Schriml and colleagues from the Disease Ontology for OMIM mappings and Sirarat

Sarntivijai, Oliver He, Alexander Diehl and Terry Meehan for discussions on the cell line model. **Funding:** The European Molecular Biology Laboratory, and EC (HEALTH theme no. 200754 Gen2Phen).

References

1. The OBI Consortium (2010) Modeling experimental processes with OBI. *J. Biomedical Semantics*, 1(Suppl 1):S7.
2. Malone, J. Holloway, E. Adamusiak, T. Zheng, J. Kolesnikov, N. Zhukova, A. Kapushesky, M. Brazma, A. Parkinson, H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112-1118.
3. Mungall, C. Gkoutos, G. Smith, C. Haendel, M. and Lewis, S. and Ashburner, M. (2010) Integrating phenotype ontologies across multiple species. *Genome Biology* ((1))R2.
4. The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25-9.
5. Falconer, S. Noy, N. and Storey, MA. (2007): Ontology mapping – a user survey. In Shvaiko, P. Euzenat, J. Giunchiglia, F. and He, B. eds.: *Proc. of the Workshop on Ontology Matching (OM2007) at ISWC/ASWC2007*, Busan, South Korea.
6. Antezana, E. et al. (2009) The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. *Genome Biology*, 10(5):R58.
7. Rector, AL. (2003) Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. *Proc. of 2nd Int. Conf. on Knowledge Capture 2003*.
8. Motik, B. Shearer, R. and Horrocks, I. (2009) Hypertableau reasoning for description logics. *J. of Artificial Intelligence Research*, 36:165–228.
9. Sarntivijai, S. et al. (2011) Cell Line Ontology: Redesigning Cell Line Knowledgebase to Aid Integrative Translational Informatics. *ICBO 2011*, Buffalo. *Accepted*.
10. Meehan, TF. Masci, AM. Abdulla, A. Cowell, LG. Blake, JA. Mungall, CJ. Diehl, AD. (2011) Logical development of the cell ontology. *BMC Bioinformatics*, 12:6.
11. Dahdul, WM. et al. (2010) The Teleost Anatomy Ontology: Anatomical representation for the genomics age. *Systematic Biol.* 59(4): 369–383.
12. Travillian, RS. Adamusiak, T. Burdett, T. Gruenberger, M. Hancock, J. Mallon, A-M. Malone, J. Schofield, P. and Parkinson, H. (2010) Anatomy ontologies and potential users: Bridging the gap. *Proc. of the Workshop on Ontologies in Biomedicine and Life Sciences*, Mannheim, Germany, 2010.
13. Bard, JBL. Malone, J. Rayner, TF. and Parkinson, H. (2008) Minimal anatomy terminology (MAT): a species-independent terminology for anatomical mapping and retrieval. In *Proc. of ISMB 2008 SIG meeting on Bio-ontologies*, Toronto.
14. OMIM, Online Mendelian Inheritance in Man (2011) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University and National Center for Biotechnology Information, National Library of Medicine, (date accessed: January 12, 2011). URL: <http://www.ncbi.nlm.nih.gov/omim/>
15. Noy, NF. et al. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* 1;37(Web Server issue):W170-3.
16. Sayers, EW. et al. (2009) Database resources of the National Center for Biotechnology Information. *Nuc. Aci. Res.* 37(Database issue):D5-15.
17. Horridge, M. Drummond, N. Goodwin, J. Rector, A. Stevens, R. and Wang, H. (2006): The Manchester OWL syntax. In *OWLed 2006*.
18. Smith, B. et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25:1251-1255.
19. Stevens, R. Malone, J. Williams, S. Power, R. and Third, A. (2011) Automating generation of textual class definitions from OWL to English. *J. Biomedical Semantics*, 2(suppl 2):S5.

Cell Line Ontology: Redesigning the Cell Line Knowledgebase to Aid Integrative Translational Informatics

Sirarat Sarntivijai¹, Zuoshuang Xiang¹, Terrence F. Meehan², Alexander D. Diehl³, Uma Vempati⁴,
Stephan Schurer⁴, Chao Pang⁵, James Malone⁵, Helen Parkinson⁵, Brian D. Athey¹, Yongqun He¹

¹University of Michigan Medical School, Ann Arbor, MI, USA

²Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA

³University at Buffalo School of Medicine and Biomedical Sciences, Buffalo, NY, USA

⁴Miller School of Medicine, University of Miami, FL, USA

⁵EMBL-EBI European Bioinformatics Institute, Hinxton, UK

Abstract. The Cell Line Ontology (CLO) is a community-based ontology in the domain of biological cell lines with a focus on permanent cell lines from culture collections. Upper ontology structures that frame the skeleton of CLO include the Basic Formal Ontology and Relation Ontology. Cell lines contained in CLO are associated with terms from other ontologies such as Cell Type Ontology, NCBI Taxonomy, and Ontology for Biomedical Investigation. A common design pattern for the cell line is used to model cell lines and their attributes, with the Jurkat cell line as an example. Currently CLO contains over 36,000 cell line entries obtained from ATCC, HyperCLDB, Coriell, and by manual curation. The cell lines are derived from 194 cell types, 656 anatomical entries, and 217 organisms. The OWL-based CLO is machine-readable and can be used in various applications.

Keywords: Cell line, Cell Line Ontology, CLO

1 Introduction

A cell line is a colony of cells that is artificially developed and grown under controlled conditions. A cell line typically derives from a multicellular eukaryote. A cell line may derive from a ‘normal’ or modified/disease tissue. A cell line can be maintained as a stable *permanent* cell lineage for renewable usages, or it may be used as a *primary* cell lineage without a long-term maintenance.

Cell lines have been widely used in research. Information about cell lines is stored in public repositories and/or indexed catalogues available for open access, and cell lines are commercially available or they are transferred between academic laboratories. Information about cell lines has not been well standardized and machine-readable to date. Each commercial provider generates a catalogue, and academic cell lines are not necessarily included. Integration of data from multiple sources is confounded by: lack of consistent naming conventions for cell lines across providers, contamination of cell lines as they are passaged and transferred between

laboratories, and provision of the same cell lines by multiple commercial sources but with different biological attributes. To address these issues, we previously produced a normalized catalogue of the Cell Line Knowledgebase (CLKB; <http://clkb.ncibi.org/>) as a project in the National Center for Integrative Biomedical Informatics (NCIBI) [1]. Since the release of CLKB, biomedical research has rapidly evolved toward integrative translational bioinformatics. In order to support translational research, conform to OBO foundry standards, and produce a resource that can be used in queries and data integration we have transformed the CLKB into an ontology available in OWL format (<http://www.w3.org/TR/owl-guide/>). Here we present the design patterns, design methodology, and content of the Cell Line Ontology – CLO.

When the Cell Type Ontology (CL) was first introduced to represent *in vivo* cell types [2], primary and permanent cell lines were included in the ontology and no separate cell line ontology existed. The Cell Type Ontology no longer includes primary or permanent cell

lines as the CLO has now become the source ontology for permanent cell lines as agreed by the maintainers of the CL, the Ontology for Biomedical Investigations (OBI), and OBO Foundry. The top-level terminology required for generating a primary cell line is provided by the OBI. The CLO is therefore a collaborative development between the CL, OBI and the CLKB developers at NCIBI and references terms from these and other ontologies in the definitions and modeling of cell lines.

In addition to CLO design and methodology, we also include examples and applications of the CLO in this study.

2 Method

2.1 Cell Line Data Sources for CLO Development

The CLO uses data from multiple sources, which are described in Table 1. The CLO cell line data were first drawn from CLKB entries, which consist of 8740 cell lines stored in ATCC (<http://www.atcc.org/>) and HyperCLDB (<http://bioinformatics.istge.it/cldb/>). CLKB will be kept as a backup source but will become obsolete at the release of the new CLO. Additional 27,000 permanent cell lines are obtained from European Bioinformatics Institute Coriell Catalogue Ontology that models cell lines from the Coriell cell repository (<http://ccr.coriell.org/>), and cell lines (both primary and permanent) provided by the Bioassay Ontology (BAO; <http://bioassayontology.org/>) development team. Cell lines that are listed in multiple repositories contain cross-reference pointers to these repositories. Cell line names can be misleading. Similar or synonymous names do not guarantee identical cell lines. Automatic mapping and manual annotation have been combined to ensure correct cell line annotation in CLO.

2.2 Importing External Ontology Terms by OntoFox

CLO imports the whole Basic Formal Ontology (BFO) [3] as its upper level ontology and the Relation Ontology (RO) [4] as its core relations. The use of these ontologies promotes integration as these resources are used by many biomedical ontologies. We used OntoFox

[5] - a technology for merging ontologies to integrate external ontologies such as NCBI_Taxon and Cell Type Ontology into the CLO. All namespaces are preserved for these ontology terms.

2.3 Definition and Annotation of CLO-Specific Ontology Terms

All cell lines and cell line-specific terms are given unified CLO IDs. The cell line data from the Coriell Cell Line ontology (<http://bioportal.bioontology.org/ontologies/45331>) have been merged to CLO with newly assigned CLO IDs. The BioAssay Ontology (BAO) has also provided a list of cell lines for inclusion in CLO. In these two cases a namespace is not preserved. When a cell line term is imported from the Coriell Cell Line ontology, we have provided a cross reference to the ontology using the *seeAlso* annotation property. Using the annotation property *comment*, BAO is noted as the source for those cell lines coming from BAO. A cell line design pattern is developed to make generic pattern between CLO cell lines and other ontology terms.

2.4 CLO Editing and Access

The development of CLO follows the OBO Foundry principles [6]. Specifically, we use unique IDs, and provide text definition for each cell line. The Web Ontology Language (OWL) is used as the default CLO format. CLO is edited using Protégé 4 Ontology Editor (<http://protege.stanford.edu>). The latest CLO is available for public view and download at

<http://sourceforge.net/projects/clo-ontology/>.

The latest version of CLO is also available for visualization and download from NCBO BioPortal:

<http://purl.bioontology.org/ontology/CLO>.

3 Results

3.1 CLO Top Structure and Statistics

The key top level classes in CLO are shown in Fig. 1. To support data integration and automated reasoning, CLO imports many terms from existing ontologies as upper level terms (e.g., *material_entity* from BFO) or terms needed for association (e.g., *cell* in CL). Cell line-specific terms are assigned with CLO

IDs (Fig. 1). The CLO-specific class *cell line* is the parent class for all specific cell lines in the CLO. The classes *permanent cell line* and *primary cell line* are the major differentia based on culture for cell lines in the CLO at present. The majority of cell lines in the CLO are permanent cell lines. Normalized cell line entries are entered as asserted CLO classes under these two subclasses. A cell line can be cultured or modified, and supplied or managed

by a cell line repository (e.g., ATCC) (Fig. 1). The detailed relations among these terms are described in our cell line design pattern (Fig. 2).

Currently CLO contains 8797 cell line-specific terms with unique CLO identifiers. In total, CLO contains 38172 terms (Table 1). The Coriell cell line records were integrated and assigned CLO-specific identifiers.

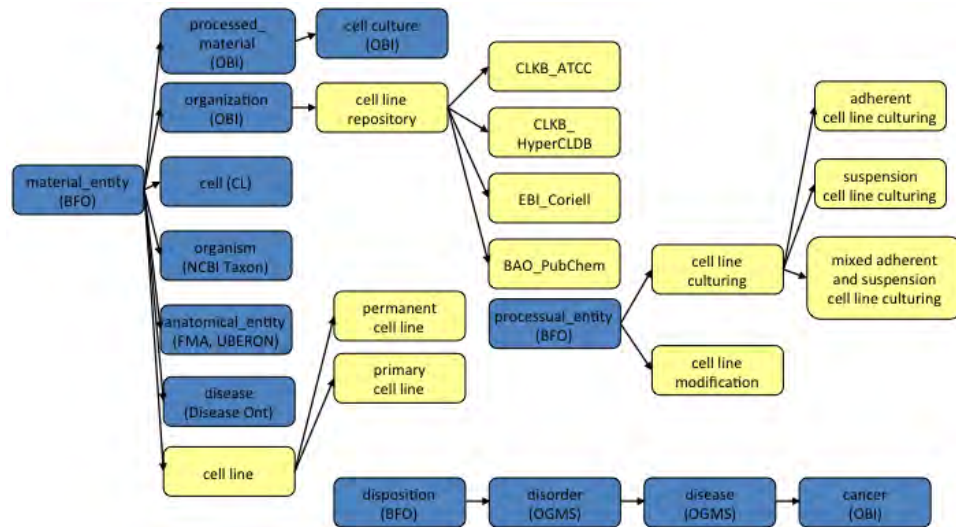


Figure 1. The top level CLO hierarchical structure of ontology terms. The terms in light blue boxes are imported from existing ontologies. The terms shown in light yellow boxes are terms with CLO unique IDs.

Ontology	Classes	Object Properties	Datatype Properties	Total
CLO (Cell Line Ontology) specific	36879	14	0	36893
Imported full ontologies				
BFO (Basic Formal Ontology)	39	0	0	39
RO (Relation Ontology)	6	25	0	31
IAO (Information Artifact Ontology)	102	14	5	121
Imported terms from other external ontologies				
OBI (Ontology for Biomedical Investigation)	15	6	0	21
CL (Cell Type Ontology)	194	0	0	194
UBERON	622	34	0	656
NCBITaxon (NCBI Taxonomy)	217	0	0	217
Total	38074	93	5	38172

Table 1. Summary of ontology terms in CLO and source ontologies used in CLO.

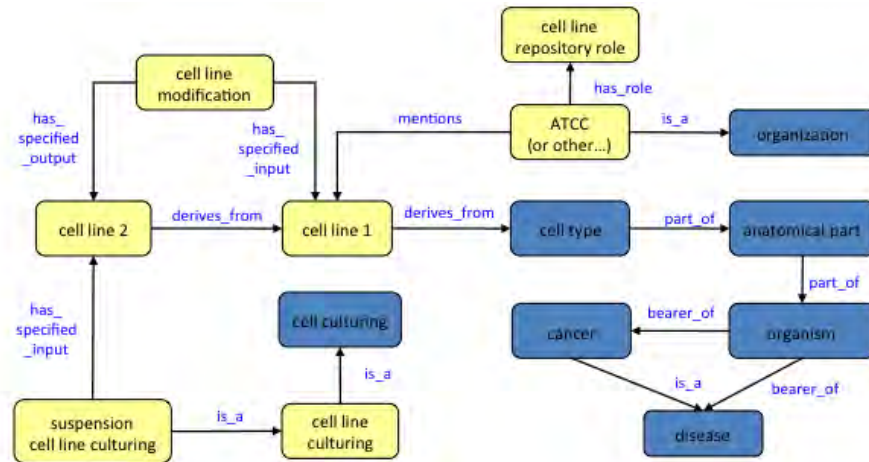


Figure 2. Basic design pattern for representing cell lines in CLO. Components shown in yellow boxes are specific to CLO, while those in blue boxes signify classes imported from other ontologies. Depending on the cell line being described, *suspension cell line culturing* and *ATCC* can be replaced with another cell line culturing process and another cell line repository, respectively.

3.2 The CLO Cell Line Design Pattern

The CLO design pattern supports representation of anatomy, cell types, disease and pathology, source information in the form of ownership and derivation where cell lines are related, and technical information such as culture conditions. Fig. 2 depicts the design pattern developed to model this information retrieved from data sources. Briefly, a cell line is originally derived from a cell type that is part of an anatomical part (e.g., liver) of a specific organism (e.g., human) having a cancer (e.g., lymphoma) or some other disease. A cell line can be derived from another cell line through a particular cell line modification. A cell line is cultured differently (e.g., *suspension cell line culturing*), which reflects a particular culturing condition or growth mode (e.g., suspension). A cell line is supplied, owned, or managed by a specific organization such as *ATCC* that has a *cell line repository role*. Since relation terms such as *supply*, *own*, or *manage* do not exist in any ontology, we use a similar relation, *mentions* (Fig. 2).

The basic cell line design pattern is followed in our CLO development. In many cases, we have also extended this design pattern by adding more content. For example, we may add a sex (female or male) quality to the organism. The pathology of the cell lines in Coriell represents one of the most important aspects to users of these artifacts and many are used as models for a particular disease. A cell line may derive from an organism that has a

specific disease (Fig. 2). The *is_model_for* relation is used to link a disease to a cell line. This relation has been created as a shortcut relation to represent the association between a cell line and a disease. In the case that a cell line derives from a normal tissue, the information of disease is omitted.

A deep understanding of the cell line design pattern that portrays a true composite architecture of cell lines requires more discussion and explanation on the relationship between CLO, CL and NCBI Taxonomy. More information is provided below.

A cell line is derived from a cell type (Fig. 2). The CL developers have been working with the CLO to ensure adequate representation of cell types from which cell lines originate. This allows mappings between the CLO and the CL using the *derives_from* relationship. Such integration also promotes error detection. To enhance interoperability with other OBO Foundry ontologies, CL-CLO mapping associates cell-types with anatomical structures using the species-neutral UBERON ontology. Thus, mappings between CLO and CL allow for associations from cell lines to anatomical structures. Sometimes a cell line cannot be mapped directly to CL as the cell line may contain multiple cell types, which can be a case of anatomical part + cell type (e.g., HCC cell line is annotated as having tissue type '*mammary gland, epithelial*'), or cell type + pathological description (e.g., AtT-20 cell line is annotated as having tissue type '*pituitary tumor, small, rounded*'), or multiple cell types

(e.g., p53NiS1 cell line has annotated tissue type *fibrous histiocytoma, fibroblast*). In this case, this cell line is related to all associated cell types using the same *derives_from* relation. According to the original repository, a cell line may derive from a cell type named by its associated anatomical part (e.g., *liver cell, peripheral blood*). These anatomy associated cell type terms have been added to CL as new CL terms to support this design. In total, 194 CL terms and 656 UBERON terms are imported to CLO (Table 1) and CLO development has expanded the CL.

The NCBI Taxonomy is the source ontology for CLO to import organism information associated with individual cell lines (Table 1). A cell line may be listed as a hybrid from multiple organisms and therefore organism and not species is modeled. In this situation, the cell line will be linked to multiple organisms. One exception of this mapping occurs when a cell line is recorded as being part of mouse/rat hybrid as there exists a class named *Mus musculus x Rattus norvegicus* as a special class of the taxonomy. Investigation of the NCBI Taxonomy also reveals that a few classes relating to those of a cell line have a place holder within NCBI Taxon, such as *mouse/rat hybrid cell lines being classified under parent term unclassified Muridae*. However, since the primary purpose of importing NCBI Taxon terms to CLO is to use the information to define organism classes, and not to redefine cell lines, we do not import these terms to CLO. A few organism values that could not be mapped to NCBI Taxon appeared to be the result of typographical errors or spelling variants. For example, there are a few cell line entries with annotated organism ‘*Agrothis segetum*’, which is believed to be a spelling variation of ‘*Agrotis segetum*’ (NCBITaxon: 47767). We do not omit or modify these original values, keeping ‘*Agrothis segetum*’ as obtained from the source (e.g., ATCC), and putting a remark in the cell line class’ comment with the information pointing to NCBITaxon: 47767.

3.3 Describing Cell Line with CLO: The Example of Jurkat

We have modeled the Jurkat Clone E6-1 cell line (ATCC # TIB-152) and its derived cell line J.CaM1.6 (ATCC #CRL-2063) as a

demonstration of our cell line design pattern usage (Fig. 3). Jurkat Clone E6-1 is a clone of an immortalized line (Jurkat) of T lymphocyte cells that was established in the late 1970s from the peripheral blood of a 14 year old boy with T cell leukemia ([7]). The J.CaM1.6 cell line is a derivative mutant of Jurkat E6-1 by treatment with ethylmethanesulfonate (EMS). J.CaM1.6 cells are deficient in Lck kinase activity and miss exon 7 in their lck mRNA.

It is noted that J.CaM1.6 cell line is not a child term of Jurkat Clone E6-1 cell line in CLO. Cell lines derived from one base cell line (e.g., J.CaM1.6 cell line deriving from Jurkat Clone E6-1) is by definition not an *is_a* relation to the base cell line but rather a *derives_from* relation. Based on this *derives_from* relation, we generate a term *Jurkat derivative cell line*, and J.CaM1.6 cell line can be inferred to be a *Jurkat derivative cell line*.

The sharing of tissue, tumor, and organism can be used to group different cell lines, such as Jurkat and Jurkat Clone E6-1. The original value ‘*peripheral blood*’ obtained from source is mapped to anatomical term ‘*blood*’ that best fits this term mapping as there are no such terms in FMA or UBERON that describe peripheral blood. Furthermore, a cell line deriving from T cell such as Jurkat is potentially problematic as T cells scatter throughout the body. Not all T cells are *part_of* some blood. Jurkat was extracted from an instance of lymphoma that was in blood. But CL’s definition of lymphocytes does not restrict all lymphocytes to blood tissue. This information is however described by ‘*isolation*’ (an OBI term used to associate tissue and cell type) inside CL. A cell line’s description embedded in CLO is specific to each individual cell line being conceptualized in each class. Reasoning with knowledge obtained from CLO and CL can capture this issue of specificity.

Many CLO specific terms (e.g., *peripheral blood cell line*) have been generated. A reasoner can be used to infer what cell lines belong to such CLO terms. Such terms are needed for many applications. For example, the ArrayExpress staff needs to know all the blood-derived cell lines and cell types for a meta-analysis of gene expression data on blood. Without such defined classes, it is difficult to obtain the results.

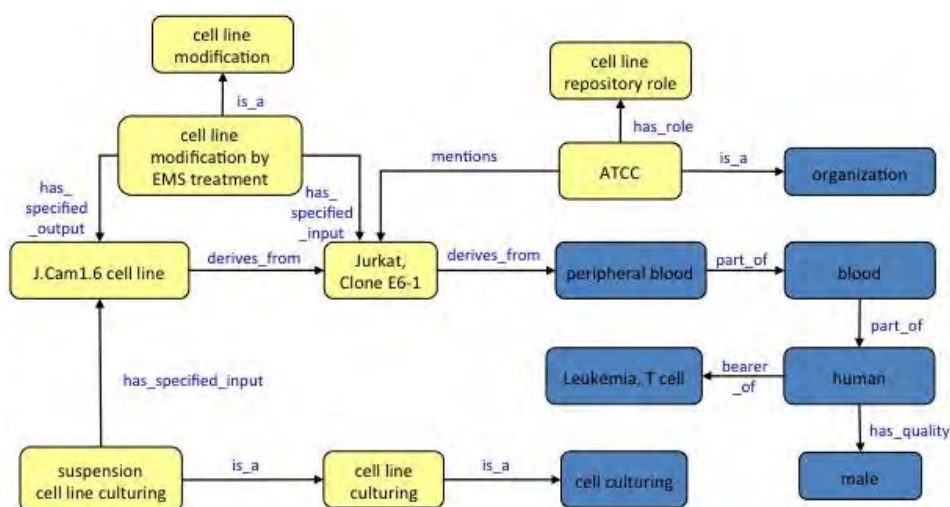


Figure 3. Modeling Jurkat and its derivative Jurkat Clone J.Cam1.6 using CLO.

3.4 CLO Application Use Case: Application of CLO in Bioassay Data Analysis

The Bioassay Ontology (BAO – <http://bioassay.ontology.org/>) describes bioassays and results obtained from small molecule perturbations, such as those in the PubChem database [8]. Integrating a formal representation of cell lines will benefit researchers in interpreting and analyzing cell-based screening results. It will also enable linking PubChem assays to other types of information (such as diseases and pathways). Moreover, formally described cell lines can help researchers in the design of novel assays, for example with respect to choosing the best cellular model system, and also in identifying which modified cell lines are available and which ones work best in existing assays.

To describe and annotate cell-based PubChem assays and screening results comprehensively, BAO is being extended through collaborative development of the CLO. By integrating BAO with CLO, those cell lines that are typically used in cellular assays are added into CLO. Based on the demands of BAO bioassay modeling, extended parameters are being added to CLO, including

different sources of cell lines (normal/healthy tissue, pathological tissue, or tumor), cell modification methods (plasmid transfection, viral transduction, cell fusion, *etc.*), culture condition (composition of culture medium), morphology (epithelial, lymphoblast, *etc.*), growth properties (adherent or suspension), short tandem repeat (STR) profiling and other properties that are relevant for cellular screening.

As a demonstration of the use of CLO in BAO bioassay modeling, we have modeled the HeLa cell line in the context of a PubChem assay (AID 1611) (Fig. 4). HeLa is an immortal cell line established from cervical adenocarcinoma of a patient in 1951 [9] and available from the ATCC (catalog # CCL-2). In the PubChem assay, HeLa cells were modified by stable transfection with a heat shock promoter driven-luciferase reporter gene construct. In this assay, the modified HeLa cells were used to screen for compounds that could induce heat shock transcriptional response as a potential therapeutic for Huntington's disease and amyotrophic lateral sclerosis (ALS).

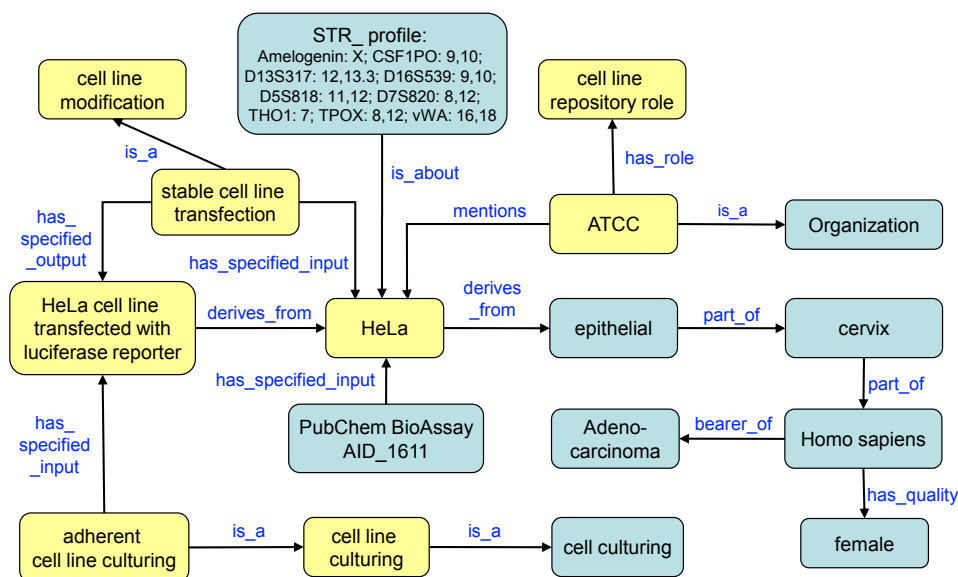


Figure 4. Application of CLO in bioassay data integration and analysis.

HeLa as demonstrated in figure 4 can be described as a *permanent cell line*, which is 'part_of' *cervix* and is 'derived_from' *Homo sapiens* that is 'bearer_of' cervical carcinoma. *HeLa* is an *epithelial cell of cervix* (CL:0002535), whose *growth mode* is *adherent*. Describing the other details of the assay are out of the scope of this paper, as they require concepts from BAO.

Many other CLO applications are being studied. For example, cell line knowledge can be used for microarray data analysis. A separate paper has been submitted to the International Conference on Biomedical Ontology (<http://icbo.buffao.edu/>) that provides more details of how the Coriell Cell Line Ontology, which has been merged to CLO, is used for ArrayExpress microarray data analysis.

4 Discussion

The availability of cellular assays and the ability to sequence DNA and RNA from single cells has promoted the use of cell lines in research and highlighted the role of cell lines in biomedical research. The release of CLO is therefore timely and will support many applications in biomedical informatics. First, CLO can be used as a tool to facilitate the data entry process for public cell line repositories (e.g., ATCC) and the referencing of these by

resources such as archival repositories (e.g., ArrayExpress). Cross-referencing with other source ontologies that are imported to CLO will allow a standard controlled vocabulary to be utilized at the data-entry point to avoid typographic errors and aid better annotation, while the depositor can also verify if the cell line being deposited already exists in the ontology, thus eliminating redundant data. Although there are currently no central authorities to assess cell line nomenclature and a cell line name is often assigned by the lab of origin, utilizing the CLO structure to frame the process will help reduce the use of duplicate names for different cell lines. It is our plan to solicit directions of new cell lines to CLO through a community-based agreement. This is also a crucial step to achieve an efficient cell line authentication process.

Furthermore, information stored in CLO can potentially validate other existing cell culture information in various sources. Gene expression data that contain the information of cell line used can be analyzed to observe if there is any data inconsistency when compared back to the information received in CLO based on the same cell line. Inconsistency in the record's attributes such as organism, tissue, tumor, or genetic mutations based on the cell line's modification may signify the possibility of cross contamination.

Cell line contamination occurs easily. It is

reported that 15% of the times cell lines being used are not what they are assumed to be [10]. Contamination also leads to issue of misidentification and mislabelling. To address this issue of contamination and mislabelling and improve cell line authentication, the American Type Cell Culture: Standards Development Organization (ATCC SDO) has proposed to establish a community-supported central authority and to use short tandem repeats (STR) as one method of verification. As a normalized indexed catalogue with ontological structure and semantics, the CLO will play a critical role in standardizing and representing cell lines and properly addressing the issue of cell line contamination. CLO can also be further expanded to link out to this STR verification information of each cell line. CLO is currently being studied for use in the ATCC SDO's authentication process.

Normalized cell line data and additional features in CLO also support applications in translational informatics such as cell line-disease association analysis, annotations of complex organ/tissue in cell cultures, and combined studies of cell culture and bioassay data.

The creation of international BioSamples databases at the EBI (<http://www.ebi.ac.uk/biosamples>) and NCBI (<http://www.ncbi.nlm.nih.gov/biosample>) provides a strong use case in that storage of non-standardized data on thousands of cell lines is not useful for high level query purposes and queries such as 'retrieve data on all ENCODE cell lines' or 'all Drosophila cell lines' will be facilitated by the addition of defined classes to the CLO, and the submission process to archival repositories will be easier if the users are able to query using the CLO ontology to retrieve validated cell line information instead of providing all this information again.

Future work of the CLO development includes the insertion of more cell lines and cell line-associated attributes. Additional CLO applications are under investigation.

Acknowledgments

We acknowledge and appreciate the following support: NIH grants 1R01AI081062, U54-DA-021519 for the National Center for Integrative Biomedical Informatics (NCIBI), NHGRI

ARRA Administrative grant HG002273-09Z (CL), RC2 HG005668 (BAO), and Gen2Phen EMBL contract number 200754 (EBI).

References

1. Sarntivijai, S., Ade, A.S., Athey, B.D., States, D.J.: A Bioinformatics analysis of the cell line nomenclature. *J. Bioinformatics* 24(23), 2760--2766 (2008)
2. Bard, J., Rhee, S.Y., Ashburner, M.: An ontology for cell types. *Genome Biol.* 6, R21 (2005)
3. Arp, R., Smith, B. Function, Role, and Disposition in Basic Formal Ontology. *Nat. Preced.* hdl:10101/npre.2008.1941.1 (2008)
4. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. *Genome Biol.* 6(5): R46 (2005)
5. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. *BMC Res. Notes* 22(3), 175 (2010)
6. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J.; OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25(11), 1251-1255 (2007)
7. Weiss, A., Wiskocil, R.L., Stobo, J.D.: The role of T3 surface molecules in the activation of human T cells; a two-stimulus requirement for IL2 productions reflects events occurring at a pre-translational level. *J. Immunol.* 133(1), 123-128 (1984)
8. Schürer, S.C., Vempati, U., Smith, R., Southern, M., and Lemmon, V.: BioAssay Ontology Annotations Facilitate Cross-Analysis of Diverse High-Throughput Screening Data Sets. *J. Biomol. Screen* 16, 415-426 (2011)
9. Scherer, W.F., Syverton, J.T., Gey, G.O.: Studies on propagations in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. *J. Exp. Med.* 97(5), 695-710 (1953)
10. Drexler, H.G., Quentmeier, H., Dirks, W.G., Uphoff, C.C., MacLeod, R.A.: DNA profiling and cytogenetic analysis of cell line WSU-CLL reveal cross-contamination with cell line REH (pre B-ALL). *Leukemia* 16(9), 1868-1870 (2002)

Towards a Semantic Web Application: Ontology-Driven Ortholog Clustering Analysis

Yu Lin, Zuoshuang Xiang, Yongqun He

Center for Computational Medicine and Bioinformatics, Unit of Laboratory Animal Medicine, and
Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, USA

Abstract. Ontology is the foundation of Semantic Web applications. The Clusters of Orthologous Groups (COG) system uses evolutionary relationships to cluster proteins from different genomes into different functional categories. In this study, we generated a COG Analysis Ontology (CAO), and used it to develop OntoCOG, an ontology-based Semantic Web application for COG-based gene set enrichment analysis. As a use case, OntoCOG is applied to a list of *B. melitensis* virulence factors retrieved from the Brucellosis Ontology (BO). This OntoCOG analysis confirms and expands current knowledge about *B. melitensis* virulence factors.

Keywords: Semantic Web, Clusters of Orthologous Groups, COG, gene set enrichment analysis, *Brucella* virulence factors

1 Introduction

The Semantic Web is a group of methods and technologies designed to allow machines to understand the meaning – or “semantics” – of information on the World Wide Web (WWW). It comprises standards and tools associated with XML, XML Schema, RDF, RDF Schema and OWL and organized in the Semantic Web Stack. During the last decade the number and scope of Semantic Web applications has remarkably increased.

Ontologies are consensus-based controlled vocabularies of terms and relations with associated definitions, which are logically formulated to promote automated reasoning. In biomedicine, ontologies play important roles such as: (a) knowledge management, including the indexing and retrieval of data and information; (b) data integration, exchange and semantic interoperability; and (c) decision support and reasoning [1]. Machine-readable ontologies play a fundamental role in Semantic Web applications in ensuring that computers can understand the semantics of terms.

The relationships between genes from different genomes are naturally represented as a system of homologous families that include both orthologs and paralogs. Orthologs are genes in different species that have

evolved from a common ancestral gene by speciation [2]. Orthologs usually share the same functions in the course of evolution. Therefore, identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. The Clusters of Orthologous Groups (COG) database (<http://www.ncbi.nlm.nih.gov/COG/>) provides a system designed to classify proteins in terms of orthologous relationships based on comparative genomic study [3, 4]. The authors of COG database defined the COGs of proteins by strictly applying all against all BLAST alignments of protein sequences from completely sequenced microbial genomes [5]. Each protein in COG database has been assigned a COG ID, and further clustered into 25 COG functional categories. These 25 Functional Categories belong to four divisions, namely: *Information storage and processing*, *Cellular processes and signaling*, *Metabolism*, and *Poorly characterized*. The COG assignment thus falls into a hierarchy fashion.

Similar with the Gene Ontology (GO; <http://www.geneontology.org/>), the COG categories can be used to perform functional analysis, i.e., COG-based gene set enrichment analysis. A COG-based gene set enrichment analysis (in short, COG enrichment analysis) serves to identify COG terms that are enriched to a statistically significant degree

among a given list of proteins compared to the distribution of these terms within the organism. Specifically, given a list of k COG annotated proteins with a total of t proteins from one organism. For a given COG category $catA$, there are q proteins within k and m proteins within t associated with it. The data will look like this in a 2×2 table:

	Given list	Not given list	Total
$catA$	q	$m-q$	M
non- $catA$	$k-q$	$t-m-(k-q)$	$t-m$
total	K	$t-k$	T

The COG enrichment analysis is to find out the statistical significance of the distribution of the data, particularly, the p-value to test whether COG category $catA$ annotated protein q is enriched (unevenly distributed) among the given protein list t . A statistical method, for example, Fisher's exact test, Chi squared test, or hyper geometric test, can be used depending on the sample size of the dataset.

There is no platform independent COG enrichment analysis package available yet, here we introduce a Semantic Web application OntoCOG, which is an ontology-oriented COG-based gene enrichment analysis. OntoCOG has a simple interface for scientists to process their data and return a result of COG enrichment analysis in OWL format. The COG enrichment analysis of *Brucella* virulence factors is used as an example to demonstrate the OntoCOG system, including the COG Analysis Ontology (CAO), an essential part for the OntoCOG design and construction. This project provides a clear demonstration on how an important biomedical question can be addressed using ontology-based Semantic Web technology.

2 Methods

2.1 OntoCOG Design and System Architecture

OntoCOG is designed as a Semantic Web service application for COG enrichment analysis. The OntoCOG software takes a given list of protein identifiers as input, performs statistical COG enrichment analysis, and returns COG analysis results as output using RDF/XML format that is modeled by CAO. In

OntoCOG, the data obtained from the COG database is stored locally in an OntoCOG relational database management system (RDMS). Each time a user sends requirements through the interface; OntoCOG will retrieve the COG annotation from RDMS and then transform the data set into a RDF/XML file. An OWL reasoner will then be applied to check the consistency and if needed remove duplicated or invalid data. Meanwhile, statistical calculation for the COG enrichment measurement will be performed, and the result is transformed into RDF based on the COG Analysis Ontology (CAO). Output data is also available in plain text format (Fig. 1).

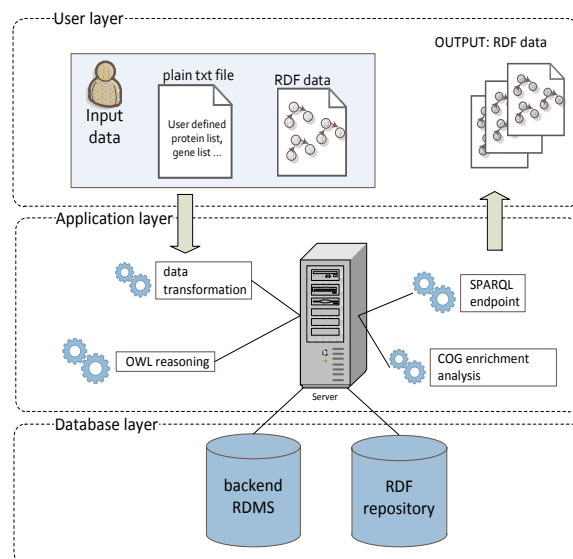


Figure 1. Design & Architecture of OntoCOG.

A conventional three-tier architecture is used for the OntoCOG system development (Fig. 1). For the user layer, a user presents data (plain text data or RDF data) or analysis queries using a front-end web browser via an HTML form or by uploading input data as a file through the interface. The middle tier, also called the application layer, extracts the input data from the user layer and implements the application's functionality. Basic functions in OntoCOG include: data transformation, OWL reasoning, COG enrichment analysis, and SPARQL data retrieval. These processes are executed with PHP scripts against the OntoCOG relational database and any publicly available RDF repository (back-end, database server). The result of each query is presented to a user

through the web browser using HTML and RDF format.

2.2 Development of the COG Analysis Ontology (CAO)

The CAO is developed based on the Semantic Web application's needs. The scopes of CAO include: 1) ontology-based software/service design; 2) supporting data integration and exchange in OWL format. The domains covered by CAO include statistical analysis and protein's COG annotation. OWL is the default format for CAO development. CAO was edited using Protégé 4.1 Beta (build 218) as ontology editor. CAO fully imports the Basic Formal Ontology (BFO; <http://www.ifomis.org/bfo/>) as its top ontology and the Relation Ontology (RO; <http://www.obo.foundry.org/ro/>) as a collective of core relations. OntoFox, an ontology development tool for importing external terms from existing ontologies [6], was used to import the following groups of ontology terms: (a) statistical analysis related terms, such as use curly quotes Fisher's exact test and Chi square test, from the Ontology for Biomedical

Investigation (OBI) [7]; (b) informatics related terms, such as data item and data set, from the Information Artifact Ontology (IAO) [8]; and (c) Organism terms from NCBITaxon [9].

3 Results

3.1 COG Analysis Ontology (CAO)

The CAO source code is available in sourceforge (<http://cao.svn.sourceforge.net/>). All the classes in CAO fall into three top classes: 1) *data transformation*, subclass of *planned process*; 2) *material entity*, subclass of *independent continuant*; and 3) *information content entity*, subclass of *generically dependent continuant* (Fig. 2).

The version 1.0 of CAO contains 178 CAO specific terms. CAO supports the design of the OntoCOG web services in aspects of data input and output flow, data modeling, and logic processing of the whole system (Fig. 3). The ontology term *OntoCOG*, asserted as a subclass of *software* (IAO_0000010), has *COG enrichment analysis data transformation objective* as its objective specification part.

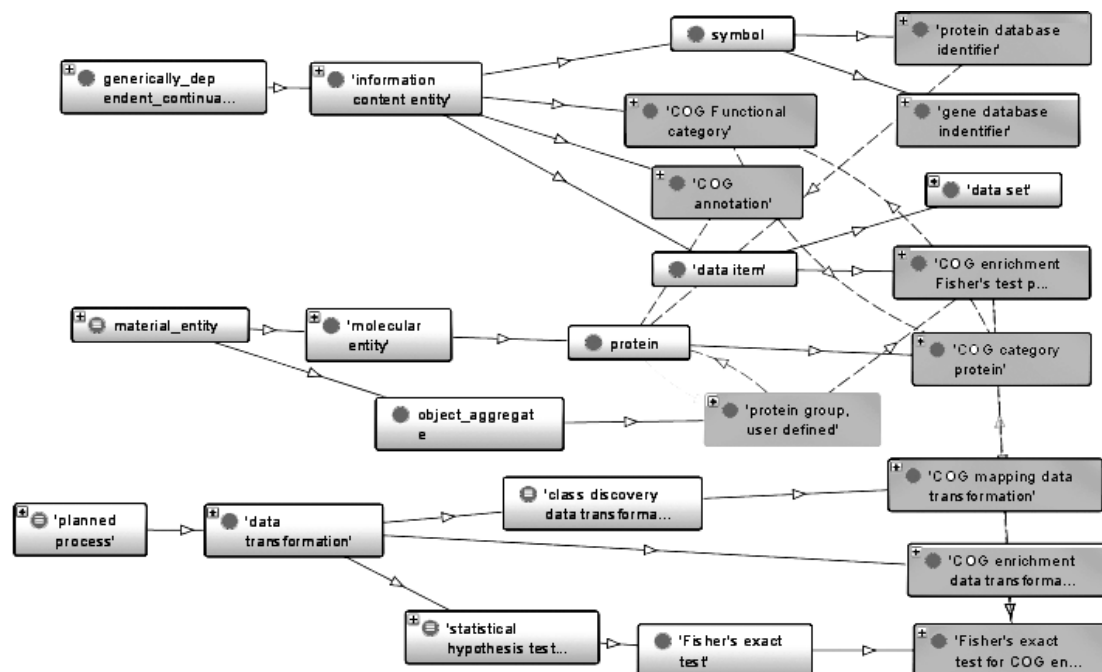


Figure 2. Key terms in the CAO hierarchy.

Gray boxes contain specific CAO terms. The remaining boxes contain terms derived from external ontologies.

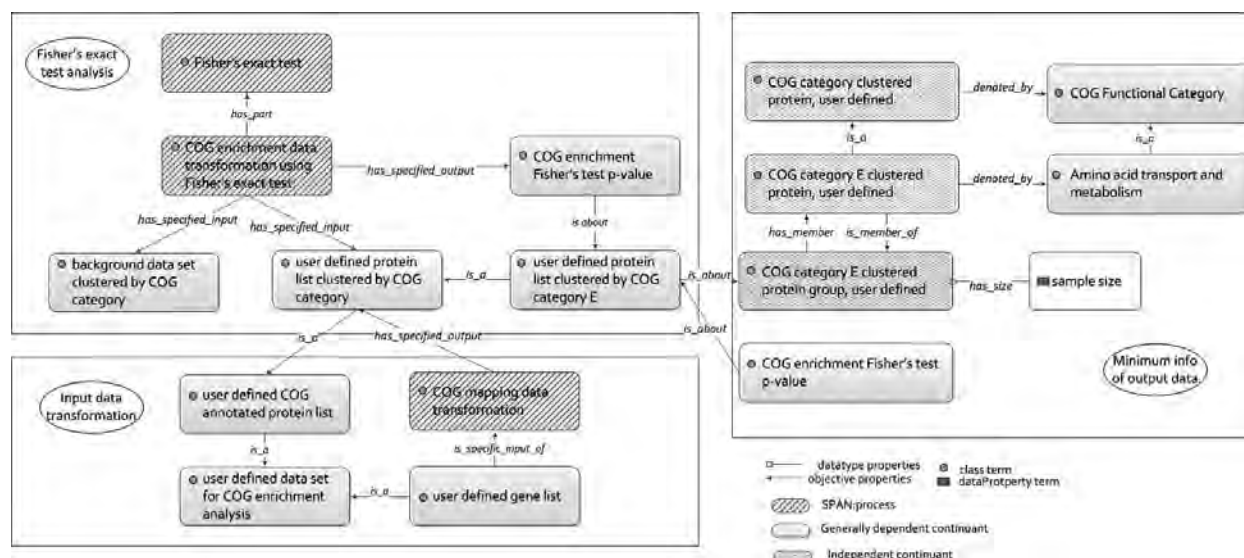


Figure 3. Design of CAO under the scope of the OntoCOG application.

CAO includes models for major components of the OntoCOG application: input data transformation, Fisher's exact Test analysis, and minimum information of output data. Terms in boxes with lines, white boxes, and boxes with dots denote *processes*, *generally dependent continuants*, and *independent continuants*, respectively.

Modeling OntoCOG input data transformation in CAO. The input dataset of OntoCOG is a list of proteins or gene identifiers submitted by a user. CAO uses IAO term *symbol* to represent the protein identifiers and gene identifiers. Both *gene database identifier* and *protein database identifier* are subclasses of *symbol*. *NCBI protein GI* and *NCBI protein accession* are subclasses of *protein database identifier*. The IAO relation *is about* has been used here to denote the relation between *information content entity* and the *material entity*. For example, a *NCBI protein GI* is about a *protein* that is a *material entity*.

By default, the input of OntoCOG is a list of NCBI protein GIs. Users may also submit a list of NCBI protein accessions or NCBI gene GIs. The server of OntoCOG will map those lists to their NCBI protein GIs, and then map the protein list with the backend COG database installed in the server. CAO models this process as a *COG mapping data transformation* process, which has a *user-defined protein* (or *user defined gene list*) as a specific input, and a protein list assigned by COG categories as one of the two specific outputs. Another specific output of this process is the sub list of proteins grouped by each COG category.

Modeling OntoCOG Fisher's exact analysis in CAO. In the CAO ontology, *COG enrichment data transformation using Fisher's exact test* is a subclass of *data transformation* (OBI_0200000). Two specified inputs are participants of this process: *background data set clustered by COG category* and *user defined protein sub list clustered by COG category*. Background data set is all the proteins assigned with COG Categories from the same species. This data set has been preinstalled into OntoCOG server as a copy of the COG database.

User defined protein list clustered by COG category is an *information content entity*, and it *is about* the material entity's aggregate: *user defined protein group*. In the following use case, an example of user defined protein group includes all the proteins annotated by one COG category.

A Perl library

Text::NSP::Measures::2D::Fisher::twotailed

that runs COG enrichment analysis by Fisher's exact test has been asserted as a subclass of *algorithm* (IAO_0000064).

The specified output of *COG enrichment data transformation using Fisher's exact test* is *COG enrichment Fisher's test p-value*.

Modeling OntoCOG output data in CAO. CAO captures the minimum information for a COG enrichment analysis. CAO specifies relevant COG categories of proteins and a p-value that explains the significance of the distribution of the list of input proteins compared to that of the whole protein list in the same organism.

In the bioinformatics field, “proteins” are often treated as data, or the system ignores the reality of a protein as a material entity. However, in ontology, “protein as material entity” and “protein as data” are distinct from each other. Recognizing this distinction will avoid the vagueness and inconsistency found in many Semantic Web applications. While a protein molecule is a material entity, a protein list is a type of datum. In CAO, several terms such as *user defined COG annotated protein list* are generated to represent data set instead of material entity.

In CAO, *protein* and *protein group* represent the major subtypes of *material entity* in the ontology. If a *protein* has been assigned by a COG functional category, it will be classified as a *COG functional category clustered protein*. For example, a *COG category E clustered protein, user defined* infers that this protein has been assigned for COG category E: *amino acid transport and metabolism*, and is a member of *user-defined COG category E clustered protein group* (Fig. 2). The size of this group is represented as a datatype property of the group. Both *user-defined COG category E clustered protein list* (a subclass of *clustered dataset*) and *COG enrichment Fishers test p-value* are information entity about this group of protein.

CAO includes several specific objective properties. Three CAO-specific relations have been created. The term *denoted_by* describes a relation between an independent entity and a data item. The domain of this property is *information content entity*, and the range is *independent entity*. Its range and domain are opposite to those of *is about*. However, *denoted_by* is not the inverse property of *is about* because there exists a many-to-many relation between an entity and

its associated information. Examples of the usage of this new relation in CAO includes: a *COG category protein* is denoted by some *COG functional category*, and a *protein* is denoted by some *COG functional category*.

The OntoCOG relations *has_member* and *is_member_of* are a pair of inverse properties. They describe the relations of a collective of entities (*object_aggregate*) and the entities within this collectivity. Both collective and individual entities are independent continuants. Both *has_member* and *is_member_of* are relations at the instance level, meaning that all the entities within one collective must be one kind. For example, an instance of the *COG category E clustered protein group* has and only has members from all the instances of class *COG category E clustered protein, user defined*.

3.2 Validation of CAO

CAO was validated by inputting real data as instances in CAO using Protege 4. The OWL reasoner HemiT1.3.3 (<http://hermit-reasoner.com/>) was used to check the consistency and axioms defined in CAO.

Two types of data were used for the validation of CAO: 1) a list of protein identifiers followed by COG functional category annotations; 2) a list of COG category clustered protein groups followed by pre-calculated COG enrichment p-value.

Data consistency checking:

In CAO, a protein assigned by a COG functional category is represented as following: *protein17987454 denoted_by E*, where *E* is an instance of COG *amino acid transport and metabolism* (i.e., COG category E).

Reasoning experiments were performed to classify individual proteins into different classes: *COG category protein* and its subclasses. The term *protein17987454* will be inferred as an instance of *COG category E clustered protein, user defined*.

Axiom validation of CAO:

Three axioms have been validated in CAO by using a reasoner to perform the classification of input data (Fig. 4):

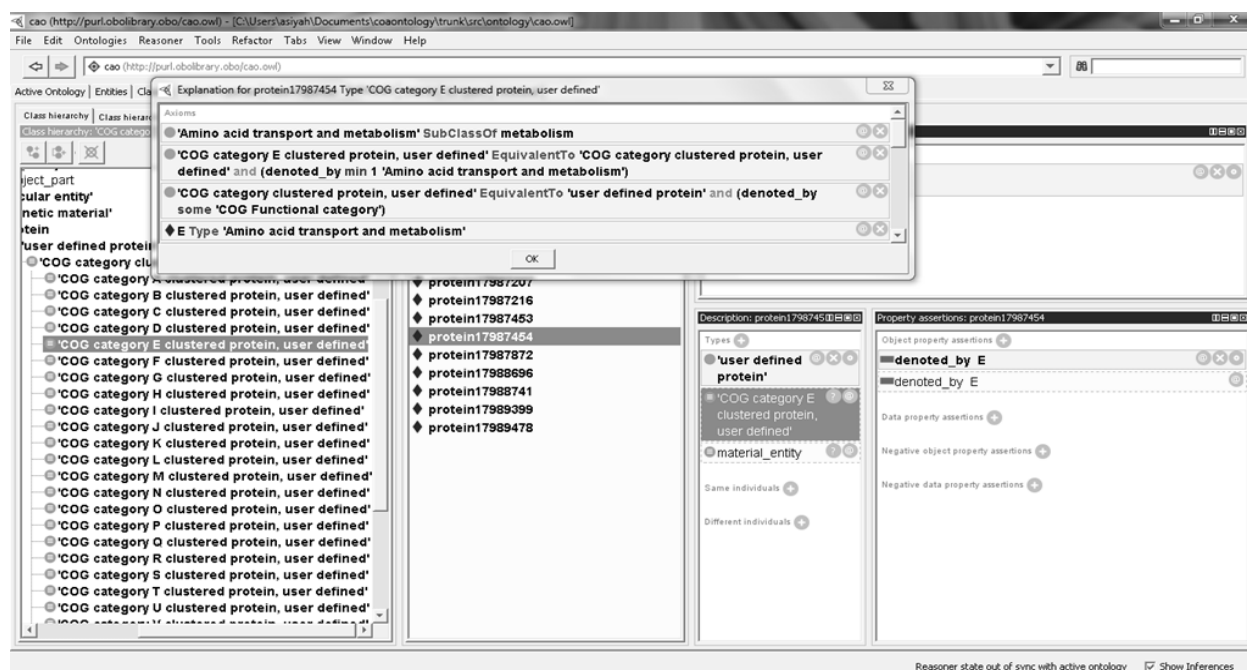


Figure 4. Automatic classification by reasoning.

A protein17987454 has been annotated as *E*, an instance of *COG amino acid transport and metabolism* (COG category E)).

By Axiom 2, this protein is classified as an instance of *COG category E clustered protein, user defined*.

Axiom 1: a *COG category clustered protein, userdefined* is a protein that has been annotated by a COG functional category in COG database:

COG category clustered protein, user defined \equiv *user defined protein* and (*denoted_by* some *COG Functional category*)

Axiom 2: a *COG category E clustered protein, user defined* is a protein from the given list that has at least 1 annotation of *COG Amino acid transport and metabolism*

COG category E clustered protein, user defined \equiv *COG category protein* and (*denoted_by* min 1 *COG Amino acid transport and metabolism*)

Axiom 3: a *COG category E clustered protein group, user defined* is a group of proteins that includes only the instances of

COG category E clustered protein, user defined
COG category E clustered protein group, user defined \equiv *protein group* and (*has_member* only *COG category E protein*)

Our studies found that all axioms and constraints in CAO are effective and efficient for data consistency checking.

3.3 Testing OntoCOG with *Brucella* Protein Virulence Factors

A list of 209 protein virulence factors from *Brucella melitensis* was obtained from the Brucellosis Ontology (BO) [10], and was submitted to OntoCOG via web interface. The COG enrichment analysis result returned by OntoCOG is shown in Table 1.

COG Category	Proteins	Fisher's exact test p-value
S: Function unknown	3	7.706e-07*
F: Nucleotide transport and metabolism	14	0.005*
R: General function prediction only	14	0.006*
N: Cell motility	8	0.011*
J: Translation	5	0.012*
G: Carbohydrate transport and metabolism	24	0.028*
U: Intracellular trafficking and secretion	9	0.052
I: Lipid transport and metabolism	3	0.083
T: Signal transduction mechanisms	11	0.111
Q: Secondary metabolites biosynthesis, transport and catabolism	1	0.126
K: Transcription	20	0.162
L: Replication, recombination and repair	6	0.237
H: Coenzyme transport and metabolism	7	0.323
O: Posttranslational modification, protein turnover, chaperones	14	0.325
C: Energy production and conversion	11	0.395
P: Inorganic ion transport and metabolism	10	0.456
E: Amino acid transport and metabolism	32	0.463
V: Defense mechanisms	3	1.000
M: Cell wall/membrane biogenesis	14	1.000
* Statistically significant (p<0.05)		

Table 1. The COG enrichment analysis of 209 *B. melitensis* virulence factors

The OntoCOG analysis identified six COG categories significantly enriched (p-value < 0.05). In total, 38 *B. melitensis* virulence factors were found to play an important role in transport and metabolism of various metabolites, including nucleotides, carbohydrates, lipids, and amino acids. Many virulence factors are components of cell motility, intracellular trafficking and secretion. These results are consistent with previous reports [11], and the p-value reports provide new statistical support. The output data can be downloaded as an RDF/OWL file that uses the CAO ontology as import ontology. The following is one synapse from the output file:

```
<!--
http://purl.obolibrary.org/obo/CAO_cog_group_J -->
  <owl:NamedIndividual
rdf:about="&obo;CAO_cog_group_J">
  <rdf:type
rdf:resource="&obo;CAO_0000309"/>
    <rdfs:label>J clustered protein group, user
defined</rdfs:label>
    <obo:CAO_0000051
rdf:datatype="&xsd:int">5</obo:CAO_0000051>
    <obo:CAO_0000117
rdf:resource="&obo;CAO_fisher_test_p_value_J"/>
    <obo:CAO_0000052
rdf:resource="&obo;CAO_protein_17986440"/>
    <obo:CAO_0000052
```

```

rdf:resource="&obo;CAO_protein_17986559"/>
  <obo:CAO_0000052
rdf:resource="&obo;CAO_protein_17986763"/>
  <obo:CAO_0000052
rdf:resource="&obo;CAO_protein_17986899"/>
  <obo:CAO_0000052
rdf:resource="&obo;CAO_protein_17988198"/>
  </owl:NamedIndividual>
<!--
http://purl.obolibrary.org/obo/CAO_fisher_test_p_value_J -->
  <owl:NamedIndividual
rdf:about="&obo;CAO_fisher_test_p_value_J">
  <rdf:type
rdf:resource="&obo;CAO_0000040"/>

<rdfs:label>0.0115540005944313</rdfs:label>
  </owl:NamedIndividual>
```

4 Discussion

Both GO [12] and COG provide gene function annotation and classification. However, only a few of prokaryotic and eukaryotic species, such as *Schizosaccharomyces pombe* (fission yeast), *Saccharomyces cerevisiae* (baker's yeast) and *E. coli*, have both COG and GO annotations. In *Brucella*, only one gene BMEI0467 in *B. melitensis* has been annotated in GO. On the contrary, COG includes annotation of all genes in *Brucella*

melitensis and many other bacteria. Many existing web services (e.g., DAVID and GOEAST) can be used for GO enrichment analysis. However, no web service for COG enrichment analysis exists yet. OntoCOG is the first web application for COG enrichment analysis.

Furthermore, OntoCOG is developed as an ontology-based Semantic Web application. OntoCOG provides CAO-based RDF/XML output data, which is more expressive and more flexible in terms of data integration. For example, users can export the p-value and the list of categories according to the enrichment measurement as other web service did. The users can also explore the attributes of specific members of each category from the given list. The use of the RDF/XML data format also allows flexibility in visualization of the data.

Future work on CAO and OntoCOG includes: 1) CAO and web interface development to allow multiple types of data input, data query, and result retrieval. 2) Provide additional statistics calculations other than Fisher's exact test. 3) Development of more advanced visualization features.

Acknowledgments

This project is supported by NIH grant 1R01AI081062. We gratefully acknowledge the critical review and editing of this manuscript by Dr. Barry Smith at the State University of New York at Buffalo.

References

1. Bodenreider O: Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform* 2008;67-79.
2. Fitch WM: Distinguishing homologous from analogous proteins. *Syst Zool* 1970, 19(2):99-113.
3. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. *Science* 1997, 278(5338):631-637.
4. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003, 4:41.
5. Kaufmann M: The role of the COG database in comparative and functional genomics. *Curr Bioinform* 2006, 1(3):291-300.
6. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y: OntoFox: web-based support for ontology reuse. *BMC Res Notes* 2010, 3:175.
7. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P *et al*: Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 2010, 1 Suppl 1:S7.
8. IAO ontology: <http://code.google.com/p/information-artifact-ontology/>
9. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S *et al*: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2009, 37(Database issue):D5-15.
10. Brucellosis Ontology (BO) <http://sourceforge.net/projects/bo-ontology>.
11. Xiang Z, Zheng W, He Y: BBP: Brucella genome annotation with literature mining and curation. *BMC Bioinformatics* 2006, 7:347.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25(1):25-29.

HeartCyc, a Cardiac Cycle Process Ontology Based in the Ontology of Physics for Biology

Daniel L. Cook^{1,2,3}, Michal Galdzicki³, Maxwell L. Neal³,
Jose L.V. Mejino², John H. Gennari³

¹Department of Physiology & Biophysics

²Department of Biological Structure

³Division of Biomedical and Health Informatics

School of Medicine, University of Washington, Seattle, WA, USA

Abstract. The computational representation of biological process knowledge is fundamental to post-genomic biomedical research and clinical practice; yet there is little agreement as to how to define and classify such processes. Here we offer a physics-based ontological schema for defining and encoding biological processes in terms of the temporal intervals during which they occur and the physical properties of participating physical entities. We develop and illustrate the use of the Ontology of Physics for Biology framework by encoding the HeartCyc ontology that represents the cardiac cycle as a multiscale use-case that spans multiple biophysical domains. We discuss the significance of our physics-based approach for rigorously defining biological processes and for bridging the disparate fields of biomedical ontology and biosimulation.

Keywords: Ontology, biophysics, thermodynamics, processes, cardiac cycle

1 Introduction

Biomedical research and clinical practice depend on measuring and describing biological processes, normal and abnormal, that occur in human and non-human organisms. In the post-genomic era, the computational encoding and sharing of such knowledge increasingly depends on controlled vocabularies and biomedical ontologies as resources for defined terms and computational models. Whereas the physical participants in biological processes (genes, proteins, cell types, organs, etc.) are encoded and cataloged in a growing collection of terminologies (e.g., ChEBI [1], UMLS [2], SNOMED-CT [3]) and biomedical ontologies (e.g., FMA [4], GALEN [5], Gene Ontology (GO) [6]), the biological processes themselves are only scantily and informally represented.

For example, Gene Ontology's (GO) Biological Process ontology defines "heart

contraction" (GO:0060047) as the "...process in which the heart decreases in volume in a characteristic way to propel blood through the body." Whereas this statement describes some occurrences during cardiac contraction, it lacks the detail, structure, and rigor of process knowledge formally encoded in quantitative mathematical models of cardiac mechanics and blood flow (e.g., [7-10]). Our goal has been to develop semantics and ontological methods by which the biophysical content of such models—the physical participants, underlying biophysics, and process knowledge—can be made available in logical form. To develop and test these methods, we have heretofore, focused on annotating and merging biosimulation models [9, 11, 12], most recently in the context of the European Union Virtual Physiological Human project (VPH [13])—a large-scale, international effort to integrate and connect biomedical data with biosimulation models.

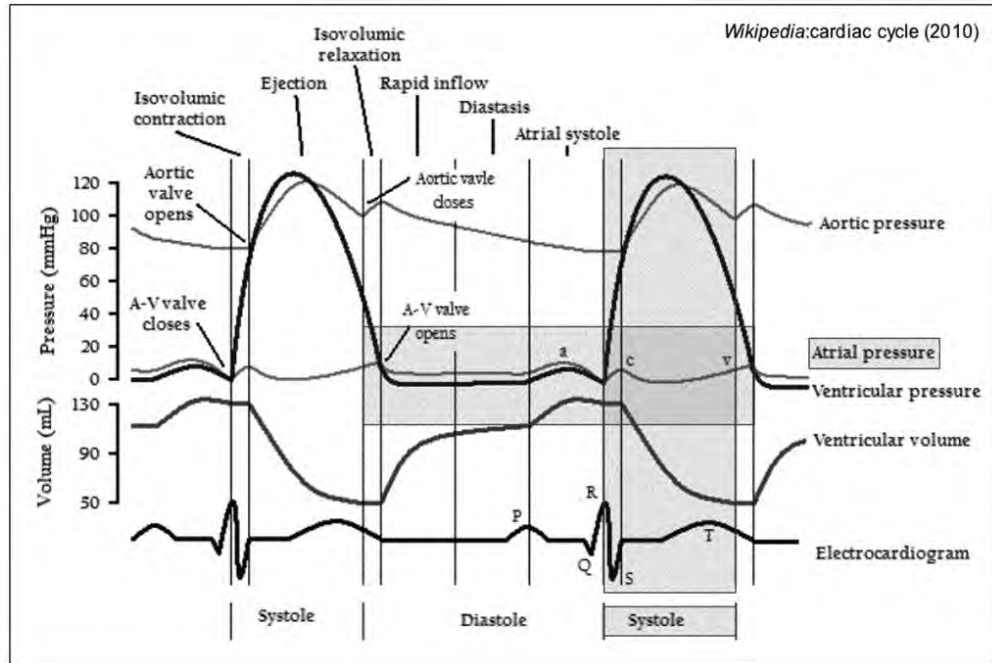


Figure 1. A Wiggers diagram graphically represents the cardiac cycle as could be recorded during a physiological experiment, as computed by a biosimulation model, or as explained to a student. Observable physical properties are labeled on the right (with calibrated scales on the left). Various temporal intervals are labeled at the bottom and top. The “cross-product” of “Atrial pressure” and “Systole” interval is high-lighted as discussed in the text.

2 Resources and Ontological Approach

2.1 Domain Knowledge: Biophysics of the Cardiac Cycle

We chose the cardiac cycle as a use-case because it is a clinically important process that has been taught to generations of physiologists and physicians and is the focus of continuing research and computational modeling in health and disease. The cardiac cycle presents a number of representational challenges as it occurs at multiple spatial and temporal scales and spans multiple biodynamic domains. For example, over the time course of seconds ventricles contract and blood flows, while on a millisecond time scale myofilaments contract triggered by fluctuations in intracellular calcium ion levels.

Although our approach to encoding biological processes is intended to generalize to all spatiotemporal scales, we here describe our HeartCyc ontology that encodes the temporal changes in physical property value during the cardiac cycle as displayed in a “classical” Wiggers diagram (Figure 1).

2.2 HeartCyc Implementation and Source Ontologies

HeartCyc is encoded in OWL¹ using the Protégé 4.1 ontology editor. We have implemented the HeartCyc ontology using classes and relations from three source ontologies:

- The Ontology of Physics for Biology (OPB [11, 14]) provides the root classes that are extended to encode HeartCyc classes that are specific to the cardiac cycle use-case. OPB classes are designated by an “opb:” prefix. HeartCyc subclasses that extend OPB carry a “heartCyc:” prefix.
- The Foundational Model of Anatomy (FMA [4, 15]) is the source of anatomical structural knowledge of the heart and those parts that participate in the cardiac cycle. FMA classes carry an “fma:” prefix.
- The Relation Ontology (RO [16]), and the

¹ Technically, there are some challenges to representing this knowledge in OWL-DL. These technical details are outside the scope of this paper, and orthogonal to our argument.

OBO Process Ontology [17], provide standardized relations such as *ro:part_of*, *ro:has_participant*, and *ro:preceded_by* to encode relations between ontology classes. Relations Ontology classes carry the “ro:” prefix.

3 HeartCyc: An Ontology of the Cardiac Cycle

To introduce the HeartCyc ontology, we describe four key aspects of our representational approach to features displayed in the Wiggers diagram. *First*, we describe observable physical properties such as “Atrial pressure”. *Second*, we describe temporal entities such as the interval “Systole”. *Third*, we describe state trajectories and state events that are cross-products of properties and temporal intervals. *Lastly*, we define dynamic processes as thermodynamic entities that are manifested by state trajectories and events. We discuss how this approach generalizes to other process domains and how it supports the temporal and structural decomposition of physical processes.

3.1 Properties of Physical Entities that Participate in the Cardiac Cycle

The first task is to encode a HeartCyc class for each of the physical properties that are named along the right hand side of the diagram. Each such property is a subclass of *OPB:Dynamical property* that is related to a physical entity by an *OPB:physical_property_of* relation. Encoding “Atrial pressure” and “Electrocardiogram” as examples, we have (shown as RDF triples):

- ```
{<heartCyc:Pressure of blood in aorta>
<rdfs:subClassOf>
 <opb:Fluid pressure>}
{<heartCyc:Pressure of blood in aorta>
<opb:physical_property_of>
 <fma:Blood in aorta>}
```
- ```
{<heartCyc:Electrocardiogram potential>
<rdfs:subClassOf>
  <opb:Electrical potential>}
{<heartCyc:Electrocardiogram potential>
<opb:physical_property_of>
  <fma:Body surface>}
```

During the cardiac cycle physical processes have cellular and molecular participants for which there is no single FMA class. For

example, a multiscale Wiggers diagram could include “Intracellular Ca^{++} concentration of ventricular myocyte” as a physical property (an *opb:Chemical concentration*) that must be composed from classes in other biomedical ontologies using RO structural relations. We encode such annotations as “composite annotations” as previously described [12].

Dynamical properties are not, of course, independent of each other. Rather, their values change according to physical laws. For example, the pressure and volume of the left ventricle depend upon each other according to a volumetric version of Hooke’s law. Aortic valve flow rate depends on the ventricular/aortic pressure difference and the valve’s fluid flow resistance according to Ohm’s law for fluids. We have developed and use semantic methods (SemGen [9, 18]) for encoding such physical properties and dependencies as semantic networks that can be decomposed and recomposed according to physical principles.

In addition to biophysical decomposition, as above, properties may be decomposed (or composed) from other properties according to structural knowledge encoded in the FMA. Thus, one could query an extended HeartCyc, for example, to discover that a property of the left ventricular wall is also a property of the heart according to the FMA’s partonomy taxonomy.

3.2 Temporal Intervals that Occur During the Cardiac Cycle

The next step for HeartCyc is to encode classes for each temporal interval (e.g., “Systole”) and each temporal instant (e.g., “Aortic valve opens”²) as labeled at the top and bottom of Figure 1. We assume that the “Systole” and “Diastole” intervals refer to processes of the left ventricle rather than of the right ventricle, or of atria (e.g., “Atrial systole”). Thus, for “Systole”, “Diastole” and the entire cardiac cycle interval (systole + diastole) we can encode a subsumption hierarchy:

² A process (e.g., valve opening) may be modeled on one time-scale as occurring in a temporal instant, yet may be modeled on another time-scale as occurring in a temporal interval.

- opb:Temporal entity
 - opb:Temporal interval
 - heartCyc:LV systole interval
 - heartCyc:LV diastole interval
 - heartCyc:Heart cycle interval

According to Figure 1, temporal intervals (opb:Temporal interval) are demarcated and bounded by temporal instants (opb:Temporal instant). For example, “Diastole” ends and “Systole” begins at the instant that the “A-V valve closes” and the “Systole” interval ends (and Diastole begins) at the instant that the “Aortic valve closes”. Thus, HeartCyc encodes heartCyc:A-V_ValveCloses instant and heartCyc:AorticValveCloses instant as subclasses of opb:Temporal instant.

The temporal relations of these intervals and instants can be encoded by structural relations to encode the part-whole relations of the intervals:

- {<heartCyc:Heart cycle interval> <ro:has_part> <heartCyc:LV systole interval>}
- {<heartCyc:Heart cycle interval> <ro:has_part> <heartCyc:LV diastole interval>}

Intervals and instants can be temporally ordered using the OPB relation opb:temporally_precedes so that, for example, the “Aortic valve closes” instant occurs as the temporal boundary between “Systole” and “Diastole”:

- {<heartCyc:LV systole interval> <opb:temporally_precedes> <heartCyc:AorticValveCloses instant>}
- {<heartCyc:AorticValveCloses instant> <opb:temporally_precedes> <heartCyc:LV diastole interval>}

The encoding of temporal instants and intervals and their mereotopological relations allows for temporal decomposition of whole processes into temporal parts. With the addition of an axiom that systole and diastole are the only parts of a cardiac cycle (i.e., a closure axiom) then HeartCyc could be queried, for example, to learn that systole and diastole are two parts of a whole cardiac cycle, that they follow each other in a cycle, and that they are separated by the closing of the aortic valve and closing of the A-V valve. This knowledge could be clinically relevant, since certain pathologies are apparent as a disordering of cardiac process

intervals and instants. For example, premature ventricular contraction (PVC) is a ventricular systole that occurs prior to atrial systole.



Figure 2. Main subclasses of OPB:Physical process entity.

3.3 Trajectories are Cross-Products of Properties and Temporal Entities

The values of a physical property (e.g., values of “Atrial pressure”) that occur during an entire cardiac cycle we define as a “state trajectory” which may be demarcated by property state events such as the occurrence of landmark property values such as a maximum value, a minimum value, or a value that traverses a designated threshold value. Figure 2 shows OPB classes that encode classes for trajectories and events. Events demarcate temporal boundaries of contiguous temporal intervals as, for example, the heartCyc:AorticValveCloses instant could be defined as the traversal of a threshold value in any of several quantitative measures of valve patency (e.g., luminal area of the valve, the pressure gradient across the valve, or the fluid flow rate through the valve).

Thus, we define opb:Property state trajectory and opb:Property value event classes as the cross-products of physical properties and temporal entities (i.e., intervals or instants). We also define a class opb:Physical event trajectory (analogous to opb:Physical state trajectory) that is an temporally-ordered aggregate of property value events that occur during a temporal interval. Physical event trajectories have their own properties such as opb:Event interval that is the duration of the temporal interval between two events in the event trajectory.

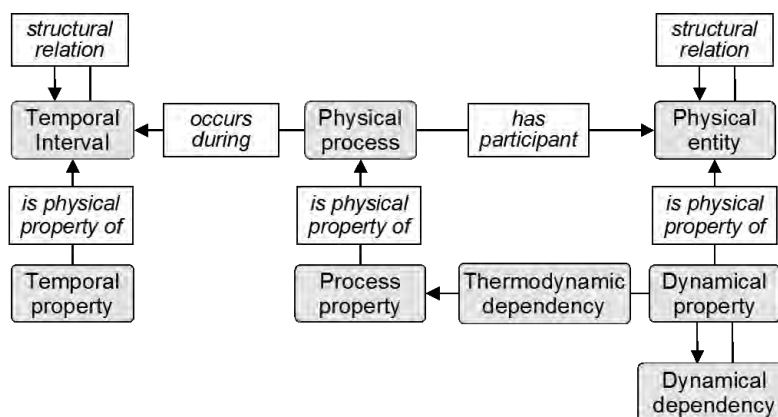


Figure 3. OPB schema for encoding biophysical process ontologies. Physical processes (opb:Physical process) have participating physical entities (opb:Physical entity) which have physical properties (opb:Dynamical property) whose values depend upon one another according to physics-based dynamical dependencies (opb:Dynamical dependency). Thermodynamic dependencies (opb:Thermodynamic dependency) define thermodynamic energy flow rates (opb:Process property) as functions of dynamical properties.

3.4 Physical Processes are Thermodynamic Entities

We have demonstrated that HeartCyc can encode the properties, intervals, and trajectories that occur during a cardiac cycle. In prior work, we have demonstrated SemSim semantic models that encode networks of physical properties linked by physics-based dependency relations. However, neither of these approaches defines physical process classes that encode, for example, how the contraction of ventricular myofibrils propels blood through the aortic valve during which molecular-level chemical processes are linked to macroscopic fluid dynamic processes.

Our hypothesis is that thermodynamics, a set of physical principles that transcend spatiotemporal scales and physical domains, provides a unifying framework for formally defining, quantifying, and encoding biological dynamical processes. Thus, we propose to define opb:Dynamical process as “...the flow, control, transformation, or dissipation of thermodynamic energy within or between participating energetic physical entities according to a physical dependency”. This view posits that the occurrence of biological processes is necessarily attended by the flow and dissipation of thermodynamic energy [19, 20] that are described by the laws and axioms of classical physics. As shown by Figure 3, these ideas are captured in the OPB schema for biological processes that relates the dynamical

properties of physical entities to process properties that are the energy flow rates that occur during a process due to changes in dynamical property values.

The cardiac cycle clearly qualifies as an opb:Dynamical process because its biological participants (the heart, its parts, and their contents) participate in an overall cardiac contractile process that includes a set of linked processes:

- the *transformation* of myocardial chemical energy into myocardial mechanical strain energy,
- the *transformation* of myocardial strain energy into fluid pressure energy of ventricular blood, and
- the *transformation* of ventricular fluid pressure energy into kinetic energy of aortic blood flow.

There are several appeals to this thermodynamic hypothesis. First, it yields to intuitive, qualitative notions of “energy” that serve well for envisioning and encoding processes yet each such process can be rigorously defined and quantitatively validated for data sets and biosimulations. Second, energy is the “common currency” of physical processes that applies as well to chemical kinetic systems as to mechanical and fluid flow systems. This offers a reduction in complexity by combining the values of domain-specific physical properties into a common quantity, energy, that reflects the thermodynamic state

of process participants. Third, because energy is a conserved quantity, one can trace the effects of one process, say myofibrillar contraction, through complex systems to determine effect on other entities, such as ventricular blood flow—“follow the energy”.

4 Summary and Discussion

We have developed and here demonstrate a prototype ontology for formally encoding dynamic biological processes as a complement to ontologies of static biological structures. We recognize that we have not evaluated our ontology but simply demonstrated how our prototype HeartCyc ontology could be used to encode key concepts in a Wiggers diagram. Thus, *opb:Physical process entity* classes can be of immediate use to the bioinformatics community for annotating, encoding, and interrelating data sets and model outputs across biomedical domains as required for multiscale integrative projects such as the VPH.

We further distinguish *properties*, *trajectories*, and *intervals* to be “manifestations” of processes rather than being processes themselves. Rather, we posit the need for a unifying theory of biological processes and propose thermodynamics as that theory. By defining physical processes as the conversion and accumulation of thermodynamic energy (for which the temporal trajectories are governed by the laws and definitions of classical physics) we can connect biomedical process ontologies to a rich legacy of prior work in physics-based mathematical models of biological processes.

References

1. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al., ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, 36(Database issue), D344-350, 2008.
2. Unified Medical Language System, <http://www.nlm.nih.gov/research/umls/>
3. SNOMED CT, <http://www.ihtsdo.org/snomed-ct/>
4. Foundational Model of Anatomy, <http://sig.biostr.washington.edu/projects/fma/>
5. Rector, A.L., Rogers, J.E., Zanstra, P.E., & Van Der Haring, E., OpenGALEN: open source medical terminology and tools. *AMIA Annu Symp Proc*, 982, 2003.
6. Lewis, S.E., Gene Ontology: looking backwards and forwards. *Genome Biol*, 6(1), 103, 2005.
7. Kerckhoffs, R.C., Neal, M.L., Gu, Q., Bassingthwaite, J.B., Omens, J.H., & McCulloch, A.D., Coupling of a 3D finite element model of cardiac ventricular mechanics to lumped systems models of the systemic and pulmonic circulation. *Ann Biomed Eng*, 35(1), 1-18, 2007.
8. McCulloch, A.D., Modeling the human cardiome in silico. *J Nucl Cardiol*, 7(5), 496-499, 2000.
9. Neal, M.L., Bassingthwaite, J.B., Subject-specific model estimation of cardiac output and blood volume during hemorrhage. *Cardiovasc Eng*, 7(3), 97-120, 2007.
10. Noble, D., Modeling the heart--from genes to cells to the whole organ. *Science*, 295(5560), 1678-1682, 2002.
11. Virtual Physiological Human Network of Excellence, <http://www.vph-noe.eu/>
12. Cook, D.L., Mejino, J.L., Neal, M.L., & Gennari, J.H., Bridging biological ontologies and biosimulation: the Ontology of Physics for Biology. *AMIA Annu Symp Proc*, 136-140, 2008.
13. Ontology of Physics for Biology, <http://bioportal.bioontology.org/ontologies/44872>
14. Foundational Model of Anatomy, <http://sig.biostr.washington.edu/projects/fma/releases/index.html>
15. Rosse, C., Mejino, J.L.V., Jr., (2007). The Foundational Model of Anatomy Ontology. In A. Burger, D. Davidson & R. Baldock (Eds.), *Anatomy Ontologies for Bioinformatics: Principles and Practice* (pp. 59-117). New York: Springer.
16. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., et al., Relations in biomedical ontologies. *Genome Biol*, 6(5), R46, 2005.
17. Ozgovde A., Gruninger M., (2010) Foundational Process Relations in Bio-Ontologies. In: Galton A, Mizoguchi R, editors. *Formal Ontology in Information Systems*: IOS Press.
18. Gennari, J.H., Neal, M.L., Galdzicki, M., & Cook, D.L., Multiple ontologies in action: Composite annotations for biosimulation models. *J Biomed Inform*, 44(1), 146-154, 2011.
19. Gennari, J.H., Neal, M.L., Carlson, B.E., & Cook, D.L., Integration of multi-scale biosimulation models via light-weight semantics. *Pac Symp Biocomput*, 414-425, 2008.
20. Neal, M.L., Gennari, J.H., Arts, T., & Cook, D.L., Advances in semantic representation for multiscale biosimulation: a case study in merging models. *Pac Symp Biocomput*, 304-315, 2009.
21. Peacocke, A.R., (1983). *An introduction to the physical chemistry of biological organization*. Oxford, UK: Clarendon Press.
22. Perelson, A.S., Network thermodynamics. An overview. *Biophys J*, 15(7), 667-685, 1975.

Composite Annotation for Heart Development

Tariq Abdulla¹, Ryan Imms¹, Jean-Marc Schleich², Ron Summers¹

¹Research School of Systems Engineering, Loughborough University, Loughborough, UK

²LTSI Signal and Image Processing Laboratory, Université de Rennes 1, Rennes France

Abstract. This paper describes progress made in combining multiple ontologies in a post-coordinated approach. This is applied to the annotation of phenotypes relevant to heart development and congenital heart diseases. The aim is to provide coordination for multiscale modeling and simulation that is presently being conducting in this field. Cardiac development is well understood within discrete levels of analysis. The application of the multiscale framework gives added value by unlocking relationships between genetic-based information at one level of analysis and the emergent phenotype at the cell and organ levels of abstraction. The challenge for a semantic representation is that this field encompasses multiple spatial and temporal scales. As a consequence, relevant terms come from a wide range of biomedical domains, and are therefore contained in several reference ontologies. The strategy for composite annotation provides a method for linking between multiscale measurement and modeling.

Keywords: Cardiac, Embryo, Morphogenesis, Imaging, Congenital, Multiscale.

1 Introduction

The explosion of data generated by many different fields of biomedical research has led to an increased focus on multiscale modeling and simulation, which provides the abstraction necessary for representing biological systems in a tractable way. The growth of computational biology is such that modeling and simulation now themselves represent a challenge in integration. The Physiome and VPH research community collaborate to provide curated model repositories of biochemical reaction networks (in SBML [1]) and biophysical mechanisms (in CellML [2]). These then have the potential to be reused in whole or part by other modelers working on different problems, potentially on different platforms, in different parts of the world. The open modeling paradigm is now so influential that some publishers advise depositing a model in these repositories as part of the publication process [3]. However, the reuse of such models will be greatly facilitated if they are represented with a common semantics, so that it is clear how the entities and parameters in one model relate to those in other models. If the same system of semantics is also used for representing experimental results, model validation will also

be greatly facilitated.

In parallel to the rise of multiscale computational modeling, a suite of reference ontologies are being developed under the umbrella of the OBO Foundry [4], which provide increasingly good coverage of biomedical concepts at different levels of spatial and temporal scale. Initially, this grew from the coordination of heterogeneous databases that record the characteristics of gene products, primarily with the Gene Ontology (GO). Reference ontologies are now used for annotating a wide variety of biomedical knowledge sources. These sources include images, database entries, publications, computational models and simulation results. By keeping reference ontologies well-bounded and essentially orthogonal the OBO Foundry minimizes logical inconsistencies and confusion over which ontology to use.

For many applications, there is a need to combine terms from multiple reference ontologies, in order to create a composite term suitable for a particular annotation. This can either be done by defining terms in application ontologies as equivalent to a composition of reference ontology terms (pre-composition); or through post-composition, whereby the annotator can compose terms on the fly, and

add them to a repository of composite terms. While the former approach is less complex for the annotator, the latter approach is more flexible.

Multiscale modeling efforts have focused mainly on the physiology of adult organ systems. Post-composed annotation of models has so far been applied only to physiological models with fairly simple physical properties [5]. The work reported here aims to tailor the multiscale framework for application to morphogenesis of the human embryonic heart. The levels of temporal and spatial scale applicable to heart development, and methods of representation are illustrated in Fig. 1. Computational modeling approaches that can be applied at different levels of scale are shown, as well as markup languages that enable a degree of model sharing between different platforms. The XML languages force a declarative expression of the components of a model, which allow it to be interpreted by different platforms. It is straightforward to annotate XML, and create an explicit link between entities in the model and external identifiers, that can be interpreted by software agents. In contrast, procedural code might be annotated with in-line comments that need a human reader to interpret them.

Along the bottom of Fig. 1, the ontologies applicable to different levels of scale are illustrated, which can be used for annotation of

different model components. These ontologies are split between occurrents, independent continuants and dependent continuants, following BFO and OBO Foundry conventions. By making this high level distinction, the OBO community has created a clearly defined boundary between the spatial and temporal domains. As simulation models comprise both domains, it is necessary to either combine terms in a post-composition approach, or make use of an application ontology for the annotation of a particular type of model.

The example process illustrated in Fig. 1 is epithelial to mesenchymal transition (EMT) in endocardial cushion growth. A signaling pathway within a single cell might be represented as an ODE within SBML. The interactions of cells and their chemical signaling might be represented with PDEs or stochastic Petri nets. A simulation of a larger numbers of cells is likely to use some form of agent based modeling. Finally, at the level of the developing heart tube as an anatomical component, finite element and multiphysics simulation may be used, to understand the relationships between mechanical properties of the heart walls (affected by EMT), its function as a pump and its looping morphology. The EMT process, and its central importance to heart development, is described in Section 2 and is used in the examples of composite annotation in Sections 3 and 4.

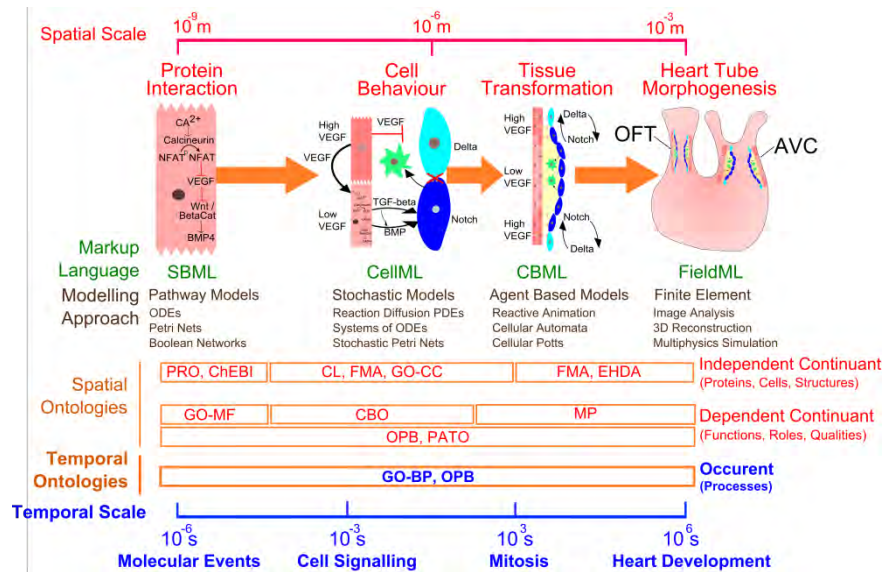


Figure 1. Spatial and temporal scales of heart morphogenesis modeling. The modeling framework encompasses spatial scales from 10^{-9} m (proteins) to 10^{-3} m (the primitive heart tube), and temporal scales from 10^{-6} s (molecular events) to 10^6 s (weeks of heart development). All acronyms are defined in the Appendix.

2 Heart Development

2.1 The Anatomic Level

The development of the embryonic heart commences in week 2 of gestation and is fully formed by week 8. This process is well documented [e.g. 6]. Week 2 of foetal life provides the first milestone of cardiac development when the two endocardial tubes that form the primitive heart fuse together. At this stage of development the first cardiac muscle contractions occur, giving rise to both blood circulation and electrophysiological signals that form a primitive electrocardiogram [7].

The embryonic heart tube is composed of an inner layer of endocardium, an outer layer of myocardium and a middle layer of extra-cellular matrix termed cardiac jelly (Fig. 2). In two restricted areas of the heart tube – the outflow tract (OFT) and atrioventricular canal (AVC) – endocardial cells adopt a mesenchymal phenotype and invade the cardiac jelly. These restricted swellings are termed ‘endocardial cushions’ and are precursors for the heart

valves and membranous septa.

The endocardial cushions begin to grow at embryonic day 26 (E26) in humans [8]. At the same time, the heart tube begins looping in an S-shape to the right. Two synchronised processes important in the understanding of congenital heart diseases are looping and aortic wedging. Looping is completed by E28, and is the first manifestation of asymmetry in the embryo. This repositioning constitutes a crucial step towards the morphology of the heart because it brings the future heart chambers and their inflow and outflow tracts into their relative spatial positions. Aortic wedging occurs as a consequence of the rotation of the myocardial wall of the OFT, itself secondary to the re-modeling of the inner curvature of the heart. Fusion of the cushions occurs at E32. In the OFT, parietal and septal cushions fuse forming the conal septum, which divides the aorta from the pulmonary artery. The conal septum is helical due to the rotation of the OFT. Upper and lower AVC cushions form the atrioventricular septum, the mitral valve and the tricuspid valve.

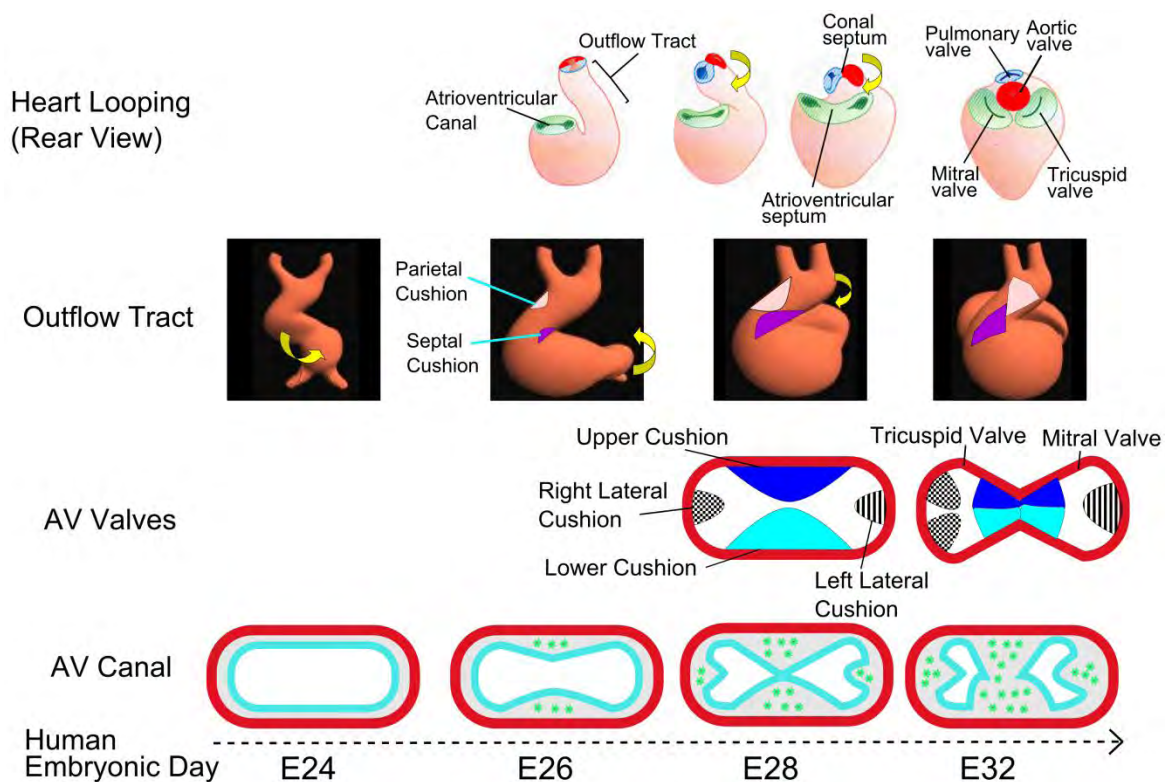


Figure 2. Detail of endocardial cushion growth and fusion. After [13].

2.2 The Cellular Level

The endocardial cushions grow by a process termed epithelial to mesenchymal transition (EMT). As the endocardial cushions play a role in forming much of the inner structure of the heart, it is apparent that abnormal EMT is a factor in many different types of congenital heart disease. These include valve, outflow tract and inter-ventricular septal defects (i.e. hole in the heart). During EMT, endothelial cells lose their adhesion to each other and invade the cardiac jelly, adopting a mesenchymal phenotype. This causes localized swelling on the inner surfaces of the embryonic heart. The study of EMT *in vitro* enables a controlled means for studying changes in cell properties, under the influence of different signaling proteins.

2.3 The Protein Level

Several signaling pathways have been identified as being important to heart development, in both the endocardium and the myocardium. Paracrine signaling acts over a short distance, perhaps between the two tissue types. These include the interactions of TGF β and BMP2 proteins, which are secreted by the myocardium in the cushion forming region. Principal among juxtacrine signaling pathways is the Notch pathway which controls pattern formation in many embryonic tissues, including those of the heart. In the endocardium, the Notch1 protein is expressed in the endocardial cushion forming regions, but not usually outside of those regions. In the myocardium the situation is reversed. Notch proteins have an additional role to play, because they activate the SNAIL family of proteins, which in turn inhibit transcription of the protein VE-Cadherin. As VE-Cadherin is one of the major proteins providing endocardial cohesion, activated Notch induces a loss of cohesion, which is part of the EMT process. It has been demonstrated *in vitro* that completion of EMT also requires BMP2 and TGF β proteins secreted by the myocardium [8]. Many types of congenital heart disease are associated with mutations in the Notch signaling pathway and this underlies the importance of Notch signaling to heart formation.

3 Methods

Developmental biology is a well established field of quantitative analysis. New results emerge every day from *in vitro* and *in vivo* high-throughput analysis, and add to the growing knowledgebase of genotype-phenotype associations. Heart morphogenesis is an area of particularly intensive research, as heart defects are among the most common type of congenital disorder. This has led to a recent expansion of the GO-BP to include a much broader range of biological process terms for heart development [9] and a corresponding initiative to increase the number of GO cardiovascular annotations. This represents a pre-composition approach, including creation of differentiation terms for 26 different cell types ('Endocardial cell differentiation', 'Pacemaker cell differentiation' etc.) Due to the logical structure of GO, these terms can be decomposed using cross-product extensions [10].

Post-composition has been applied successfully in annotating phenotypic descriptions. This makes use of a particular type of ontology composition: the Entity Quality (EQ) formalism. This extends entity terms from reference ontologies by describing them as the intersection of the entity with a relationship to a quality term in PATO. The entities are most often from species specific anatomy or developmental anatomy ontologies, but may also be a cell type from CL; a biological process, molecular function or cellular component from GO; or a molecular level entity from PRO or ChEBI. The EQ formalism has been used for investigating the evolution of phenotypic traits (phylogenetics) [11] and in integrating phenotypic annotations from multiple species [12], and in this way linking human diseases to mutant animal models [13]. In contrast, the Mammalian Phenotype (MP) ontology takes a pre-composition approach, which aims to include terms sufficient for phenotypic description within a single ontology [14]. This has been used successfully for the mouse and rat genome databases. The two approaches are not mutually exclusive, as MP terms could be defined as equivalent to EQ terms, when appropriate.

The post-composition approach has also

begun to be used for the annotation of biomedical simulation models. This is similar to the EQ formalism described above, but using the Ontology of Physics for Biology (OPB) rather than PATO. The OPB describes both physical properties and physical processes. This is because simulation models mainly represent the physical properties of biological entities. The SemGen tool enables modelers to annotate SBML or CellML code using OPB post-composition terms; although they must first be imported and compiled in the JSim modeling tool [15]. Once models are annotated in this way, a semantic comparison of several models can then be made through SemGen, automatically identifying entities that can be combined if models are merged. However, this approach to annotating models has only been applied to domains with well defined physical properties. It is not clear how well this would work for cell level modeling for example, where the physical properties that drive cell behavior are not fully understood.

It is straightforward to adapt the EQ formalism for developmental phenotypes. The initial step is to select the relevant ontologies for the domain, as well as the types of sources that might be annotated. The process for the domain of heart development is illustrated in Fig. 3.

PATO allows composite phenotype annotations such as ‘endocardial cushion with decreased concentration of SNAIL protein’, which are composed from the integration of multiple reference ontologies. OPB allows formalization of the physical properties of these composite annotations, such as the concentration of a particular protein in a particular endocardial cell, or the density of mesenchymal cells in an endocardial cushion. These terms can then be used to annotate variables in a computational model, or experimental data. PATO composites can also be mapped to disease classifications, such as OMIM.

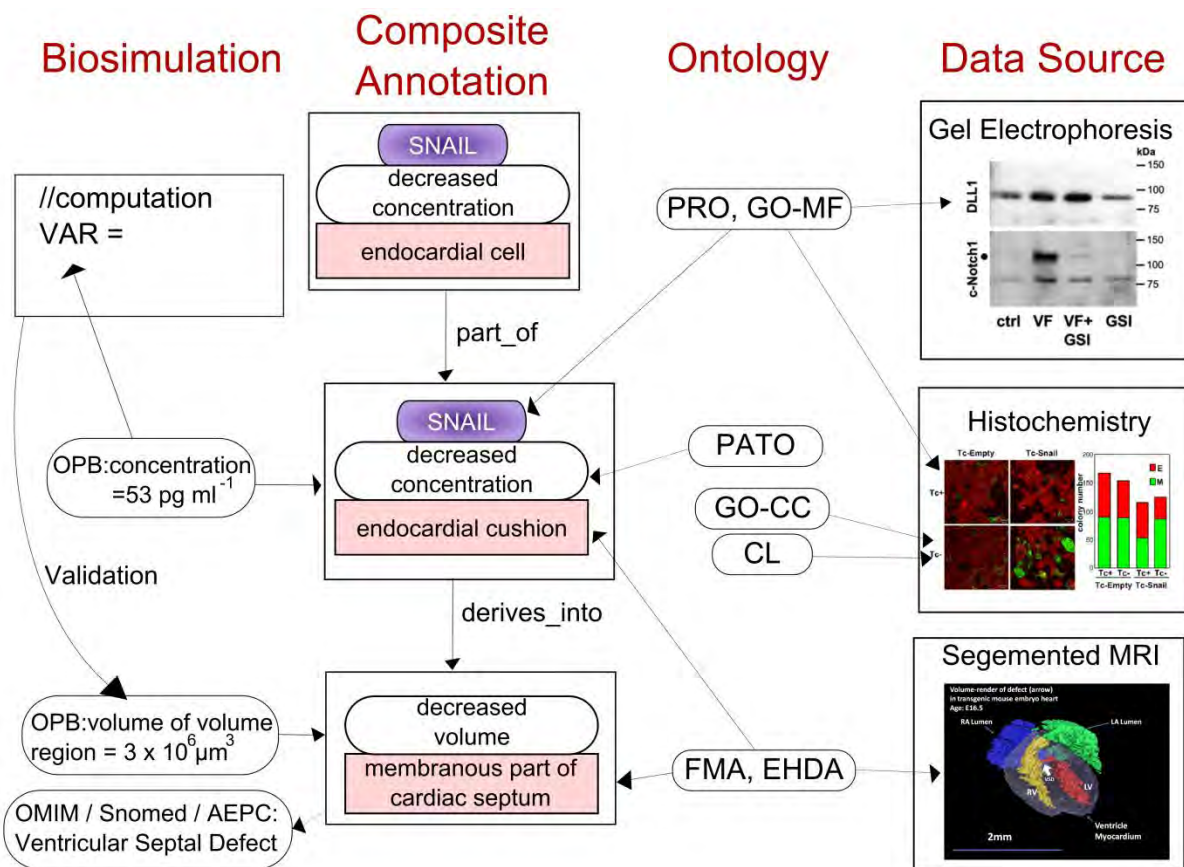


Figure 3. Schema for composite annotation. Refer to Appendix for acronyms.

OBO	OWL
intersection_of:	EquivalentTo:
PATO:0001163!decreased concentration	PATO:0001163
intersection_of: inheres_in	and (inheres_in some
PR:000015308!SNAI1	PR:000015308)
intersection_of: contained_in	and (contained_in some
CL:0002350 ! endocardial cell	CL:0002350)

Table 1. OBO and OWL representation of composite annotation.

4 Results

The topmost annotation shown in Fig. 3 can be represented in OBO or OWL format as detailed in Table 1.

Annotations will be represented using a simplified EQ syntax, using term labels rather than identification numbers. The annotation in Table 1 would be expressed as:

```
PATO:decreased concentration
<inheres_in> PR: SNAI1 <contained_in>
CL:endocardial cell
```

The OPB based composite annotation for the concentration parameter or measured value would be:

```
OPB:chemical concentration <property_of>
OPB:portion of molecules <composed_of>
PR:SNAI1 <contained_in> CL:endocardial
cell
```

The same annotation could be used whether pointing to a model parameter, or an experimentally measured concentration. This suggests a method for leveraging the semantic relationships between very different types of information.

An EQ representation may be defined under a number of categories [11], with the examples below taken from the process of heart morphogenesis.

Monadic states are those that involve single entities or structures. For example, it has been previously shown that some congenital heart abnormalities are caused by an incorrect rotation of the outflow tract. This can be annotated in a general way as:

```
PATO:mislocalised_radially<inheres_in>
EHDA:outflow_tract
```

Relational states are those that describe a phenotype that exists between two entities or structures. The first example in this section was relational.

Quantitative states describe a measured

value for a variable feature (e.g. size, area, count). For example, the volume of an endocardial cell would be annotated as:

```
OPB:volume region <inheres_in>
CL:endocardial cell <has_magnitude>
OPB:volume amount=3.2 <has_unit>
UO:microliter
```

5 Discussion

With post-composition, there is a lack of exact consistency in annotations between different annotators [12]. This is not always a major problem because, with sufficient guidelines, the differences are usually ones of specificity (e.g. did they use the FMA term ‘endothelium’, ‘endothelium of endocardium’ or ‘endothelium of aortic valve?’). These annotations are still valid semantically, but where a more coarse term is used there is a degree of information loss, to be avoided where possible. Restriction to terms of a specific domain and the use of customizable software tools for annotation improves consistency. An example of the latter is Phenote [11], an open source toolkit that facilitates annotation of biological data using OBO-format ontologies. However, it is still possible to have different perspectives on the same physiological phenomenon. For example, one decision might be whether the interest is in the decreased volume of the membranous septum, or the fact that the membranous septum is dysfunctional. From the perspective of exact volume quantification the actual size measurement is important, whereas in the more general disease classification the interest lies only in the fact that there is a dysfunction. There are often pre-composed terms in existing ontologies, which could also be made by post-composing terms from multiple ontologies. For example, in the MP ontology the term ‘abnormal outflow tract development’, could be composed as:


```
PATO:abnormal <inheres_in>  
GO:outflow_tract_morphogenesis
```

The degree of variability possible is a key advantage of post-composition: congenital heart diseases are a spectrum of overlapping phenotypes, and it is necessary to have flexibility in the way they are annotated. This accuracy in genotype-phenotype annotation, while arguably more complex, is more beneficial to wider biological research than mere coding of defects for the sake of classification. However, the strategies are not mutually exclusive. An intriguing possibility is to map anatomical measurements (such as those determined from the MRI of congenital heart disease specimens) to disease classifications.

The challenge of reasoning over multiple ontologies remains a considerable one. Nonetheless, it is much more feasible to achieve data integration in this way than in any existing alternative. In particular if new ontologies were constructed for each application, with no semantic links to existing reference ontologies, then integrating across applications would be almost as cumbersome as not using ontologies at all.

6 Conclusion

It has been argued that simple annotations (e.g. a pointer to a single reference ontology class) are insufficient for annotating the variety of data sources that need to be integrated within the current multiscale modeling projects [16]. The variety of possible classes increases due to the need for more highly specified annotations. To this extent the post-composition approach is necessary for fully integrated multiscale annotation.

Multiscale modeling research has hitherto focused almost exclusively on adult physiology, with little attention to embryonic development, although this has begun to change. The team is in the process of developing multiscale simulations of endocardial cushion growth via EMT. Cell-tissue level modeling of EMT is achieved with CompuCell3D, while signaling pathways are modeled using SBML. Combining knowledge gained from the information models (EQ formalism) allows the closure of the loop between physical experiments (real world) and computer based

simulations (model world). As the EQ annotations of the model world map to their isomorphic physical counterparts in the real world it is possible to be unambiguous about referring to (say) endocardial cells or increased concentration of a given protein.

Creating accurate phenotypic descriptions, which retain their semantic context, and linking these to physical and biophysical measurements, provides a powerful means to assimilate information from a wide variety of sources and scales. To this end the team has access to a unique physical resource – over 50 post mortem heart specimens that have been diagnosed as tetralogy of Fallot. The intention of future work is to provide MRI data of these specimens to link between the primary evidence and degree of outflow tract rotation.

A further limitation to overcome is the lack of exact consistency in ontological annotations. Nevertheless, data sources of different types, at different scales have been identified, alongside the ontologies suitable for annotation, modeling methods at different levels, and initial guidelines for composite annotation. This demonstrates a method for creating a link between multiscale measurement and multiscale modeling that assists in closing the loop between physiological and genetic understanding of cardiac development.

References

1. Novere, N.L., Courtot M., Laibe, C.: Adding Semantics in Kinetics Models of Biochemical Pathways. In: 2nd International ESCEC Symposium on Experimental Standard Conditions on Enzyme Characterizations, pp. 137–154. Beilstein-Institut, Rhein (2007)
2. Beard, D.A., Britten, R., Cooling, M.T. et al.: CellML Metadata Standards, Associated Tools and Repositories. *Phil. Trans. R. Soc. A* 367, pp. 1845–1867 (2009)
3. Courtot, L.C., Novère N.L., Laibe C.: BioModels.net Web Services, a Free and Integrated Toolkit for Computational Modeling Software. *Briefings in Bioinformatics*. 2, pp. 270–277 (2010)
4. Smith, B., Ashburner, M., Rosse, C. et al.: The OBO Foundry: Co-ordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature Biotechnology*, 25, pp. 1251–1255 (2007)
5. Neal M.L., Gennari J.H., Arts T., Cook D.L.: *Advances in Semantic Representation for*

- Multiscale Biosimulation. In: Pacific Symposium on Biocomputing, 14, pp. 304–315. WSP, Hawaii (2009)
6. Kirby, M.L.: Cardiac Development. OUP, Oxford (2007)
 7. Christoffels, V.M., Smits, G.J., Kispert, A., Moorman, A.F.M.: Development of the Pacemaker Tissues of the Heart. *Circ. Res.* 106, pp. 240–254 (2010)
 8. Luna-zurita L., Prados, B., Grego-bessa, J. et al.: Integration of a Notch-dependent Mesenchymal Gene Program and Bmp2-driven Cell Invasiveness Regulates Murine Cardiac Valve Formation. *J. Clin. Invest.*, 120, pp. 3493–3507 (2010)
 9. Khodiyar, V.K., et al.: The Representation of Heart Development in the Gene Ontology. *Developmental Biology*, 354, pp. 9-17 (2011)
 10. Mungall, C.J. et al.: Cross Product Extensions of the Gene Ontology. *Journal of Biomedical Informatics*, 44, pp. 80-86 (2011)
 11. Balhoff J.P., Dahdul, W.M., Kothari, C.R. et al.: Phenex: Ontological Annotation of Phenotypic Diversity. *PLoS ONE*, 5, e10500 (2010)
 12. Mungall, C.J., Gkoutos, G.V., Smith, C.L. et al.: Integrating Phenotype Ontologies Across Multiple Species. *Genome Biology*, 11, R2 (2010)
 13. Washington, N.L., Haendel, M.A., Mungall, C.J. et al.: Linking Human Diseases to Animal Models Using Ontology-based Phenotype Annotation. *PLoS Biology*, 7, e1000247 (2009)
 14. Smith C.L., Goldsmith C.A.W., Eppig J.T.: The Mammalian Phenotype Ontology as a Tool for Annotating, Analyzing and Comparing Phenotypic Information. *Genome Biology*, 6, R7 (2005)
 15. Gennari, J.H., Neal, M.L., Galdzicki, M., Cook, D.L.: Multiple Ontologies in Action: Composite Annotations for Biosimulation Models. *Journal of Biomedical Informatics*, 44, pp. 146–154 (2010)
 16. Cook, D.L., Mejino, J.L.V., Neal, M.L., Gennari, J.H.: Composite Annotations: Requirements for Mapping Multiscale Data and Models to Biomedical Ontologies. In: 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2791–2794. IEEE, Minnesota (2009)

Appendix: Glossary of Terms

AEPC	Association for European Paediatric Cardiology	GO-CC	Gene Ontology Cellular Component
AVC	Atrioventricular Canal	GO-MF	Gene Ontology Molecular Function
BFO	Basic Formal Ontology	HPO	Human Phenotype Ontology
CBML	Cell Behavior Markup Language	MathML	Mathematical Markup Language
CBO	Cell Behavior Ontology	MP	Mouse Phenotype
CellML	Cell Markup Language	MRI	Magnetic Resonance Imaging
CheBI	Chemical Entities of Biological Interest	OBO	Open Biomedical Ontologies
CL	Cell Type Ontology	OFT	Outflow Tract
EHDA	Edinburgh Human Developmental Anatomy	OMIM	Online Mendelian Inheritance in Man
EMT	Epithelial to Mesenchymal Transition	OPB	Ontology of Physics for Biology
EQ	Entity-Quality	PATO	Phenotype and Trait Ontology
FieldML	Field Markup Language	SBML	Systems Biology Markup Language
FMA	Foundational Model of Anatomy	VPH	Virtual Physiological Human
GO-BP	Gene Ontology Biological Process	XML	eXtensible Markup Language

Ontology-Based Analysis of Event-Related Potentials

Gwen Frishkoff^{1,2}, Robert Frank², Paea LePendou³

¹Georgia State University, Atlanta, GA, USA

²University of Oregon, Eugene, OR, USA

³Stanford University, Stanford, CA, USA

Abstract. We describe recent progress in the development and application of NEMO (Neural ElectroMagnetic Ontology), a formal ontology for the event-related potentials (ERP) domain. The ontology encodes knowledge about patterns that are commonly seen in ERP studies. The patterns are defined using equivalent class descriptions, which specify the spatial, temporal, and functional constraints that must be satisfied for an ERP instance, or datum, to belong to a particular pattern class. The data themselves are represented in RDF, using N-triples that link the data to the ontology. Our analysis pipeline automatically generates these RDF data. We then apply a reasoner, such as Hermit, to classify the data. By creating this pipeline, we have enabled our consortium partners to compare results across experiment paradigms using a common knowledge base and to refine that base (i.e., to add or adjust pattern descriptions) based on cross-lab study results. We discuss implications for ERP meta-analysis, discovery of new knowledge, and resolution of current controversies in the ERP literature.

1 Introduction

This paper describes recent progress in the development and application of NEMO (Neural ElectroMagnetic Ontology), a biomedical ontology for the event-related potentials (ERP) domain. The driving motivation for NEMO is the need to make valid comparisons across ERP datasets. Although ERPs have been used in human neuroscience for over 50 years, there have been remarkably few meta-analyses, and the few that exist are of questionable validity [1]. By contrast, although functional magnetic resonance imaging (fMRI) is a newer technique, meta-analyses are now routine in the fMRI literature [2].

One problem that hinders meta-analysis in the ERP domain is the lack of a standard vocabulary and the absence of explicit, formal definitions for patterns that are commonly seen in a particular experimental context (e.g., visual word recognition). A second challenge is the complexity of ERP data, which has led to a variety of approaches to ERP pattern extraction. These cross-lab differences may result in incommensurable data, which cannot serve as inputs to a valid meta-analysis. The goal of the NEMO project is to address these problems by developing a seamless pipeline for

ERP analysis, statistical measure generation, and classification of data, which can be used across studies and across labs.

In previous work [3-5], we described the structure of the NEMO ontology, which represents knowledge about ERPs and foundational concepts from various domains. The present paper describes how the ontology can function as a tool for classification and labeling of ERP data. Two recent developments have been central to this effort. First, we have added equivalent class descriptions (aka “rules”) for ERP pattern classes within the ontology.

Second, the instance-level data themselves are now automatically created as part of our ERP analysis pipeline. The pipeline takes raw ERP data as input, extracts pattern instances, and produces summary metrics for each pattern. The metrics are then used to generate RDF/OWL files (henceforth, “RDF data”), which contain a few basic assertions about each pattern and link them to the ontology.

As a result of formally encoding the ERP pattern classes, class descriptions (rules), and instance-level data, we can now classify real ERP data using a reasoner such as Hermit [6]. This is a major milestone for the NEMO project.

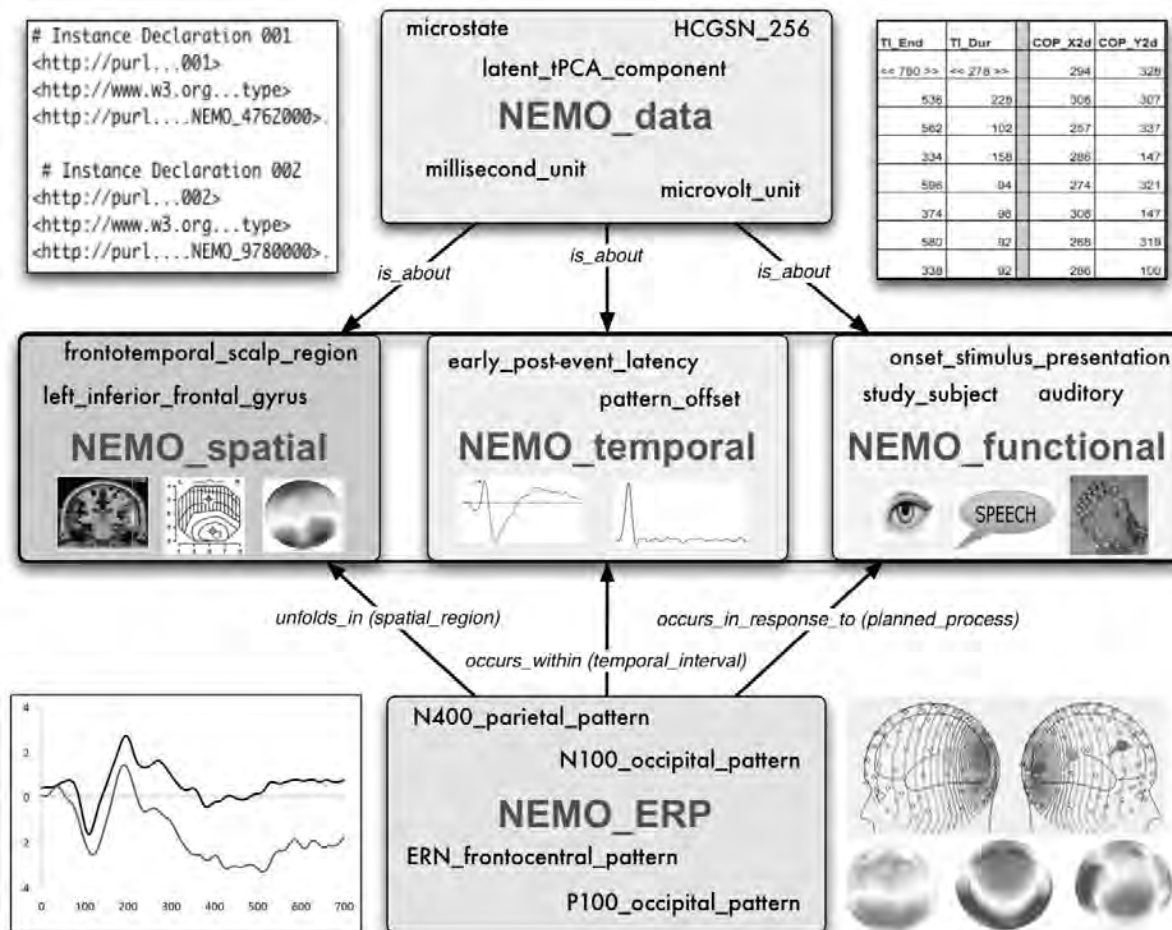


Figure 1. Core subdomains of the NEMO ontology.

Following, we outline these new developments and discuss how they can lead to scientific breakthroughs in the ERP domain. We see the development of ERP pattern classes and class descriptions (rules) as an ongoing project: researchers will ideally generate new results using the NEMO ERP analysis pipeline and refine the knowledge base to reflect new findings. We therefore emphasize the importance of considering both top-down (knowledge-driven) and bottom-up (data-driven) methods in ontology development.

2 NEMO Ontology

ERPs are measures of brain electrophysiology (“brainwaves”). ERPs provide a powerful means for studying brain function, because they are acquired noninvasively and can therefore be used in a variety of populations

(e.g., children and patients, as well as healthy adults). In addition, they provide detailed information about the time dynamics, as well as the spatial distribution, of neural activity during various cognitive and behavioral tasks.

The NEMO ontology is a domain-specific knowledge base that is built on top of the Basic Formal Ontology (BFO) [8]. As described in previous work [1, 3-5], NEMO has been designed in general to comply with OBO Foundry best practices [7]. For example, we make every effort to re-use existing ontologies. To this end, NEMO imports concepts from other ontologies, including the Ontology for Biomedical Investigations (OBI; [9]), Neuroscience Information Framework (NIF; [10-11]), the Foundational Model of Anatomy (FMA; [12]), and the Cognitive Paradigm Ontology (cogPO; [13]).

NEMO includes five core domains (see Fig.

1). The NEMO_spatial domain includes concepts representing spatial regions (e.g., brain and scalp locations) and qualities (e.g., dorsal/ventral), and anatomical entities that correspond to the locations of interest (e.g., brain, scalp, skull). NEMO_temporal comprises temporal intervals (e.g., time periods referenced to ERP experiment events, which are critical for analysis) and temporal qualities, as well as some physiological concepts. NEMO_functional includes concepts related to cognitive and behavioral processes and paradigms that are relevant during experimentation. Finally, NEMO_data includes concepts related to measurement and analysis of data (e.g., “peak latency,” “mean amplitude”). These five domains are separated only in theory. In practice, all classes and all relations are encoded in a single file.

In earlier versions we maintained separate files for each domain. However, a practical issue emerged as we started to define class restrictions: definitions often reference concepts from multiple domains. For example, NEMO defines an ERP as a type of process (NEMO_temporal), which unfolds in some spatial region (NEMO_spatial). In order to represent this information, it was therefore necessary to re-assert concepts in multiple files, to import these files (which caused performance problems), or to create bridge files [10] that would require additional work to maintain.

The latest release of the NEMO ontology can be browsed and downloaded from the

BioPortal website.

(<http://bioportal.bioontology.org/ontologies>)

All versions, including the most recent (“working”) and older (legacy) versions can be accessed from our SVN repository.

(<http://purl.bioontology.org/NEMO>)

3 ERP Pattern Rule Representation

An important goal for the NEMO ontology is to represent formally the spatiotemporal ERP patterns that have been frequently studied in cognitive and language-related ERP experiments over the past several decades. To this end, we have coded ~40 ERP pattern classes in the current version of NEMO (v. 1.60). Most recently, we have added equivalent class descriptions for each of these pattern rules. Each pattern rule specifies three sets of criteria (see Figure 2):

- (1) **Temporal.** The peak latency of a particular pattern falls within a certain time range (in milliseconds).
- (2) **Spatial.** A pattern is characterized by surface-positive and negative voltages (in microvolts), which are distributed over certain scalp regions-of-interest.
- (3) **Functional.** A particular pattern occurs within a certain experimental context, in response to specific types of experimental stimuli, response and task requirements.

```
visual_occipital_P100_pattern EquivalentTo scalp_recorded_ERP_extracted_pattern
(1) and (has_proper_part some (peak_latency_measurement_datum
    that (has_numeric_value some (decimal[>= "70"] and decimal[<= "140"]))))
(2) that ((has_proper_part some (intensity_measurement_datum
    that (is_quality_measurement_of some (intensity
        that (inheres_in some (scalp_recorded_ERP
            that (unfolds_in some occipital_scalp_surface_region))))
    and (has_numeric_value some decimal[>= ".4"^^decimal])))
(3) and (proper_part_of some (averaged_EEG_data_set
    that (is_about some (scalp_recorded_ERP
        that (occurs_in_response_to some (onset_stimulus_presentation
            that (has_object some (object
                that (has_quality some visual)
                and (has_role some stimulus_role))))))))))
```

Figure 2. Example of an ERP pattern rule.

Part (1) of this assertion expresses the **temporal criterion** for the visual_occipital_P100_pattern.

Part (2) expressed the **spatial criterion**.

Part (3) expresses the **functional (experimental) criterion**.

Figure 2 illustrates how these three criteria are used to express the pattern rule as an equivalent class description for the “visual occipital P100,” a well-known pattern in ERP research on visual perception [3-5].

4 ERP Data Representation

The NEMO ERP Analysis Toolkit is a suite of tools that provides an pipeline for ERP analysis, statistical measure generation, and creation of instance-level data that are linked to the NEMO ontology. The Toolkit uses a MATLAB class-based architecture, which allows for re-use of common objects, such as data provenance, which are referenced at every stage of the processing pipeline.

The analysis pipeline itself includes the three main steps (1-3 in Figure 3, below): (1) Step 1, ERP pattern extraction, (2) Step 2, ERP metric extraction, and (3) Step 3, RDF code generation. After initializing the script for pattern extraction (Step 1), the rest of the process is entirely automated.

Step 1: ERP Pattern Extraction. ERP pattern analysis is the process of transforming complex spatiotemporal ERP data into discrete patterns, which are used for analysis of experimental (condition) effects on the latency, amplitude, and topographic distribution of neural activity. The NEMO

toolkit includes two types of pattern analysis: Decomposition, which includes various implementations of Principal Components Analysis, or PCA, and Independent Components Analysis or ICA and Windowing, or Segmentation. In contrast with conventional methods for ERP component analysis, all of the methods in NEMO are data-driven (See Ref. [1] for details). As a result, the extraction (and subsequent definition) of a particular ERP pattern is not subject to experimenter bias. Further, data can be batch-processed for efficiency.

Step 2: ERP Metric Extraction. The ERP patterns that are extracted in Step 1 are input to the ERP Metric Extraction tool, which computes summary measures of time course (e.g., peak latency, duration) and scalp distribution (e.g., average intensity over each scalp region of interest) for each of the patterns in a particular dataset (See Ref. [1] for details).

Step 3: RDF Data Generation. Finally, the latest version of the NEMO toolkit (v. 1.18) automatically writes out the results of the metric extraction script to RDF. RDF generation is new to this project. Therefore, this process is detailed in the following section.

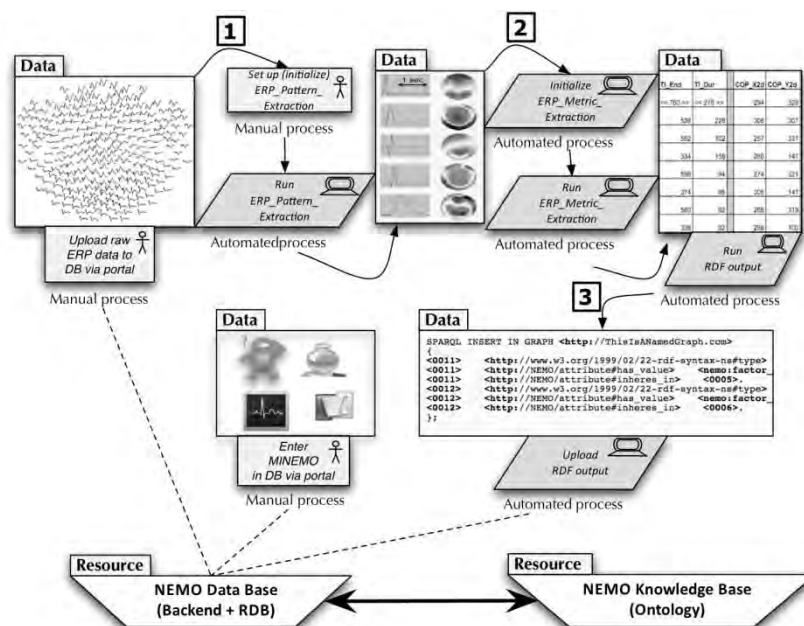


Figure 3. ERP Data processing pipeline.

The NEMO ERP Metric Extraction (Step 2 in the processing pipeline; see Fig. 2) yields a set of spatial and temporal metrics that capture the main features of ERP pattern instances. These metrics are subsequently used to classify instances, using NEMO ERP ontology rules (see following section for details on these rules).

In order to represent ERP data within the NEMO ontology, we have created a MATLAB script that writes out the summary ERP metrics to an RDF (Resource Description Framework) file. The MATLAB RDF generation treats all input/output files as distinct resources, each with a Uniform Resource Identifier reference (URIref). Similarly, each ERP data file, its attributes, the elements of its provenance, the parameters governing its transformation, the transformed data, and the file contents (ERP summary metrics) are assigned URIs that are linked to the NEMO ontology.

The MATLAB RDF runtime script assigns a set of RDF triples, or descriptive statements, to each resource. A triple consists of a *subject–predicate–object* structure, in which the predicate specifies a binary relationship between the subject and object, for example,

value001 – is_a – mean_intensity_LFRONT. All subjects, predicates and objects (except for the typed literals) are RDF resources, which are indexed by NEMO concept URIs. Thus, the RDF generation effectively “annotates” ERP data using a small set of concepts from the NEMO ontology.

RDF triples represent the minimal information that is needed to link the data to the ontology. For example, the class *mean_intensity_LFRONT* represents the average intensity over left frontal electrodes, a concrete and uncontroversial concept. Our goal was to generate data representations that are likely to be stable and uncontroversial, and are therefore unlikely to change over time. These data-related classes are linked to other parts of the NEMO ontology through a chain of assertions that are more complicated and abstract, as shown in Figure 4. Note that this more complex assertion is not part of the RDF representation. As a result, changes in scientific knowledge should not require that we re-annotate existing data. Rather, the RDF representation of the data can simply be reclassified using a new version of the ontology.

```
mean_intensity_LFRONT EquivalentTo intensity_measurement_datum
that (is_quality_measurement_of some (intensity
that (inheres_in some (scalp_recorded_ERP
that (unfolds_in some (left_frontocentral_scalp_surface_region))))))
```

Figure 4. Class restriction for mean_intensity_LFRONT (an ERP metric class).

5 Classifying and Labeling ERP Data

After a data set has been fully processed using the NEMO ERP Analysis Toolkit, the resulting data (RDF file) can be opened and processed in Protégé [15]. The first several lines of the RDF file import the NEMO ontology. Thus, the data and ontology are both available within the file. The data can then be classified using a reasoner such as Hermit [6].

Figure 5 illustrates the classification results for one such analysis. The instance-level datum ('NW_ERP_0352') is an ERP

pattern that has a pronounced surface-negativity at around 352 ms and occurs in response to a visually presneted nonword stimulus. Based on the spatial, temporal, and functional properties of this pattern, it was classified as a member of the medial_frontal_negativity (MFN).

The example in Figure 5 also illustrates an interesting scenario, which is likely to appear rather frequently in real applications: If distinct ERP pattern classes have overlapping spatial and temporal criteria, then a particular ERP observation can be classified as a member of more than one class.

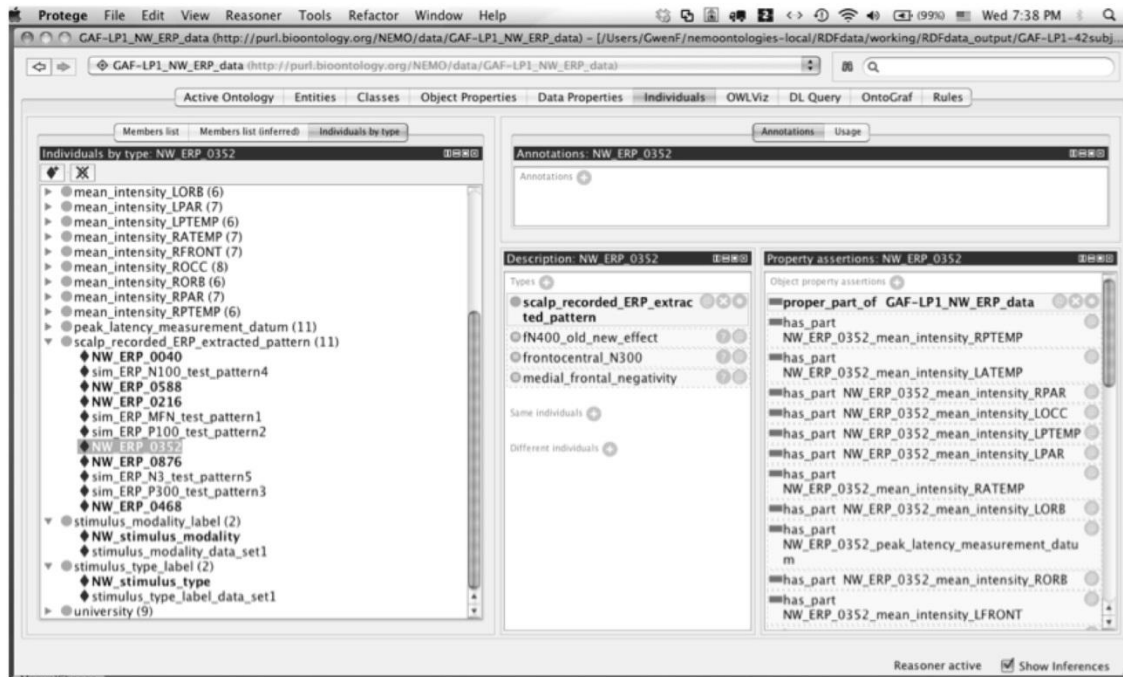


Figure 5. Classification results for one pattern (ERP response to a nonword) from a real ERP dataset.

Conversely, a pattern instance that satisfies none of the pattern rules will be classified as a member of the *undefined_ERP_pattern* class, which is defined in NEMO as the complement of all other (defined) ERP pattern classes. In each case, the classification results may challenge current definitions and, in doing so, raise a central issue for future applications: how to manage changes in the ontology over time.

To address this issue, we must first acknowledge that the most interesting parts of the ontology – that is, the pattern rules – are uncertain by their very nature. To capture this uncertainty, NEMO makes use of evidence codes, a type of annotation that has been used in GO [14] to flag the source of evidence for existence of a particular class or class definition. In NEMO, “author assertion” is considered the weakest source of evidence for a particular ERP pattern rule. The strongest evidence is a published set of results from a quantitative meta-analysis — evidence that will come with the application of NEMO tools to multiple datasets from our cross-laboratory ERP consortium.

6 Discussion

In conclusion, we have described a novel application of NEMO (Neural ElectroMagnetic Ontology), a formal ontology for the event-related potentials (ERP) domain. The ontology encodes knowledge about patterns that are commonly seen in ERP studies. The patterns are defined using equivalent class descriptions, which specify the spatial, temporal, and functional constraints that must be satisfied for an ERP instance, or datum, to belong to a particular pattern class. We have thereby attempted to capture ERP domain knowledge in a formal, explicit way. Naturally, this knowledge will evolve over time. Hence, it will be important to track and document the evidence that supports a particular pattern description and to curate this information over time.

Our hope is that this approach can help to resolve some long-standing controversies in the ERP literature. For example, the “N400” pattern has been described in more than 400 published papers, but it remains a point of controversy whether this pattern reflects automatic (e.g., unconscious) activation of word meanings or whether it is only seen in response to effortful processing of semantic information

[16]. Informally, it has been characterized as a surface-negative pattern peaking at around 400 ms over centroparietal sites [17]. However, its precise measurement and quantification vary widely, even across studies within the same research lab. This variability has made it hard even to achieve informal generalizations across ERP study results. To illustrate, Dombrowski and Heil [18] recently stated that “the interpretation of the N400 is far from being resolved.” This state-of-affairs is somewhat surprising, given 30+ years of research and several hundred publications focused on N400 semantic effects. However, the reason for this state-of-affairs is evident: there are inconsistent definitions of core concepts, such as the “N400,” in ERP research.

The ambiguity of natural-language definitions in ERP research has important implications for ERP research. In particular, it suggests that data mining from text may give unreliable results, since natural-language terms are used inconsistently and thus cannot be assumed to pick out the same real-world entities. This implies, in turn, that the inputs to ERP data mining and cross-laboratory analysis should ideally consist of either structured or semi-structured ERP data, rather than natural-language descriptions of these data. To this end, we have created a unique workflow for ERP analysis, which has several key features. First, it seamlessly combines ERP analysis, metric extraction, and RDF file generation. Thus, it fills an important gap in available tools for ERP research. Second, the workflow is fully automated, which removes the need for manual selection of spatial and temporal variables. Thus, the results of ERP analysis and metric extraction (which are inputs to pattern classification) are guaranteed to be compatible with the ontology. Third, the variables themselves are noncontroversial (measures of onset, offset, and peak latency and distribution of positive and negative potentials over different regions of the scalp). Fourth, the set of variables is extensible (for example, we are adding spectral measures to the next release), so the system can support users who wish to do something novel. Likewise, the Toolkit allows flexibility in selection of pattern extraction methods, so users are not bound to one approach. In this sense, our system is not complete, but this is

true by design: we fully expect that methods for ERP analysis will continue to evolve. Finally, the ontology and ontology-based resources for NEMO were developed in collaboration with an international group of ERP researchers, who represent different approaches and different kinds of experience with ERPs.

Our next step is to apply the NEMO ontology-based workflow to data from a variety of studies from across our 8 consortium sites. We are focusing on three related paradigms that have generally been studied in isolation from one another: (1) word and nonword recognition, (2) semantic priming, and (3) episodic memory for familiar and newly learned words. These three paradigms all evoke surface-negativities that have been related to semantic memory. Our hope is to discover similarities and differences in the brain's response to semantic memory in these different experimental contexts. This work has strong significance for reading and language development and interventions for clinical conditions, such as dyslexia and language deficits due to traumatic brain injury and stroke.

Finally, as in prior work, we emphasize the importance of both top-down (knowledge-driven) and bottom-up (data-driven) methods in ontology development [1, 3-5]. Previously, we have suggested this approach could lead to robust descriptions of ERP pattern classes. Our current approach is consistent with this top-down/bottom-up framework: whereas the initial (“seed”) versions of the ERP pattern rules are based on published literature (top-down), our approach to ERP pattern analysis is data-driven (bottom-up). The challenge is what to do when classification results suggest inconsistencies or gaps in the ontology. This question is likely to be a central topic of ongoing and future research in biomedical ontologies.

Acknowledgements

This work is supported by a grant from the National Institutes of Health, award #R01EB007684. We would also like to thank Timothy Redmond (Stanford Center for Biomedical Informatics) for helpful suggestions on coding RDF files and combining the instance-level information with the ontology.

References

1. Frishkoff G, Frank R, Rong J, Dou D, Dien J, Halderman L. A framework to support automated classification and labeling of brain electromagnetic patterns. *Comput Intell Neurosci* 2007:14567.
2. Laird, A. R., Eickhoff, S. B., Kurth, F., Fox, P. M., Uecker, A. M., Turner, J. A., et al. (2009). ALE Meta-Analysis Workflows Via the Brainmap Database: Progress Towards A Probabilistic Functional Brain Atlas. *Front Neuroinformatics*, 3, 23.
3. Dou D, Frishkoff G, Rong J, Frank R, Malony A, Tucker, D. Development of NeuroElectro-Magnetic Ontologies (NEMO): A Framework for Mining Brainwave Ontologies. Proceedings of the 13th International Conference on Knowledge Discovery & Data Mining (KDD'07) 2007:270-279.
4. Frishkoff, G. A., LePendou, P., Frank, R. M., Liu, H., & Dou, D. (2009). Development of Neural Electromagnetic Ontologies (NEMO): Ontology-based Tools for Representation and Integration of Event-related Brain Potentials. Paper presented at the *International Conference on Biomedical Ontologies (ICBO'09)*, Buffalo, NY.
5. Frishkoff, G., Sydes, J., Mueller, K., Frank, R., Curran, T., Connolly, J. F., Kilborn, K., Molfese, D., Perfetti, C. A. and Malony, A. D.: Minimal information for neural electromagnetic ontologies (minemo): A standards-compliant method for analysis and integration of event-related potentials (erp) data. In: *Standards in Genomic Sciences*, (2011, under review).
6. Shearer, R., Motik, B., & Horrocks, I. (2008). HermiT: A Highly-Efficient OWL Reasoner. Paper presented at the *OWL: Experiences and Directions (OWLED 2008)*.
7. Simon, J., Dos Santos, M., Fielding, J. and Smith, B.: Formal ontology for natural language processing and the integration of biomedical databases. In: *Int J Med Inform*, vol. 75(3-4), pp. 224-231 (2006).
8. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L. and Lewis, S.: The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. In: *Nat Biotechnol*, vol. 25(11), pp. 1251-1255 (2007).
9. Brinkman, R. R., Courtot, M., Derom, D., Fostel, J. M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Soldatova, L. N., Stoeckert, C. J., Jr., Turner, J. A. and Zheng, J.: Modeling biomedical experimental processes with obi. In: *J Biomed Semantics*, vol. 1 Suppl 1(2011, in press).
10. Bug, W. J., Ascoli, G. A., Grethe, J. S., Gupta, A., Fennema-Notestine, C., Laird, A. R., Larson, S. D., Rubin, D., Shepherd, G. M., Turner, J. A. and Martone, M. E.: The nifstd and birnlex vocabularies: Building comprehensive ontologies for neuroscience. In: *Neuroinformatics*, vol. 6(3), pp. 175-194 (2008).
11. Bug, W. J., Astahkov, V., Boline, J., Fennema-Notestine, C., Grethe, J., Gupta, A., Kennedy, D. N., Rubin, D. L., Sanders, B., Turner, J. and Martone, M.: Data federation in the biomedical informatics research network: Tools for semantic annotation and query of distributed multiscale brain data. In: *AMIA Annu Symp Proc*, vol. pp. 1220 (2008).
12. Martin, R. F., Rickard, K., Mejino, J. L., Jr., Agoncillo, A. V., Brinkley, J. F. and Rosse, C.: The evolving neuroanatomical component of the foundational model of anatomy. In: *AMIA Annu Symp Proc*, vol. pp. 927 (2003).
13. Turner, J. A. and Laird, A. R.: The cognitive paradigm ontology: Design and application. In: *Neuroinformatics*, (2011, in press).
14. Dwight, S. S., Harris, M. A., Dolinsky, K., Ball, C. A., Binkley, G., Christie, K. R., et al. (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*, 30(1), 69-72.
15. Rubin, D. L., Noy, N. F., & Musen, M. A. (2007). Protégé: a tool for managing and using terminology in radiology applications. *J Digit Imaging*, 20 Suppl 1, 34-46.
16. Mari-Beffa, P., Catena, A., Valdes, B., Cullen, D. and Houghton, G.: N400, the reference electrode, and the semantic activation in prime-task experiments: A reply to dombrowski and heil (2006). In: *Brain Res*, vol. 1147(pp. 209-212 (2007).
17. Kutas, M. and Federmeier, K. D.: Electrophysiology reveals semantic memory use in language comprehension. In: *Trends Cogn Sci*, vol. 4(12), pp. 463-470 (2000).
18. Dombrowski, J. H. and Heil, M.: Semantic activation, letter search and n400: A reply to mari-beffa, valdes, cullen, catena and houghton (2005). In: *Brain Res*, vol. 1073-1074(pp. 440-443 (2006).

River Flow Model of Diseases

Riichiro Mizoguchi¹, Kouji Kozakil¹, Hiroko Kou¹, Yuki Yamagata¹,
Takeshi Imai², Kayo Waki², Kazuhiko Ohe²

¹ISIR, Osaka University, Ibaraki, Japan

²Department of Medical Informatics, Graduate School of Medicine, The University of Tokyo, Japan

Abstract. This article discusses the ontological treatment of diseases in the framework of the Ontology for General Medical Science (OGMS). We aim to provide a definition of a disease that is more friendly to clinicians and propose a corresponding model of diseases. We define a disease as a dependent continuant constituted of one or more causal chains of clinical disorders. To clarify the ontological meaning of causal chains, we introduce two kinds of processes: a *cumulative continuous process* and a *non-cumulative process*. They are accounted for based on a new ontological theory of objects and processes. We then introduce the core ideas of a disease as causal chain and of clinical imbalance. We believe that the result can be considered as a concretization of the OGMS view of disease as disposition.

Keywords: disease ontology, causal chain, objects and processes, imbalance

1 Introduction

Recently, there has been a serious need for a consistent and ontologically sound medical vocabulary. The need is increasing as the tasks which need to be addressed by information technology in handling medical data become ever more demanding. In this situation, we believe that the Ontology for General Medical Science (OGMS) [1] based on BFO [3] and developed under the supervision of Barry Smith, is of considerable value. In OGMS, we find an excellent definition of a disease as: *a disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism*. This is a beautiful definition from a philosophical point of view. At the same time, however, it is not very friendly to clinicians because it lacks practicality. This reminds us of the reaction of engineers when they learn Smith's preferred account of function, which is also a beautiful one, as a type of disposition. However, engineers believe that function is something more real than a mere disposition. We believe that a domain ontology should be useful for both domain experts and ontologists.

We are not claiming that such a philosophically beautiful definition is useless to domain experts. Rather, we would like to try to develop another definition of disease that is more

friendly to clinicians by concretizing the notion of disposition in keeping with the philosophy of vocabulary design used by OGMS.

This paper is organized as follows. We begin by analyzing the definition of disease in OGMS and explain our motivation in developing another definition of disease. Section 3 discusses our definition of disease. Section 4 provides an ontological theory of objects and processes to support this definition. Based on our definition, we propose in Section 5 a disease model that can be implemented on a computer.

2 Analysis of the Definition of a Disease in OGMS

Our concerns about the definition of disease in OGMS are as follows:

- (1) Dispositions are introduced in the course of disease development in the human body. A disposition is a potentiality; on the OGMS view the realization of this potentiality takes the form of chains of physical/physiological changes in the human body. For OGMS currently, therefore, disease and disease course are distinguished; the latter is in a sense outside the former. We believe this use of 'disease' is counterintuitive to clinicians, and we thus propose a definition of disease that allows the disease to be placed within

the chain of events that is the disease course.

- (2) To see what is missing from the current OGMS's approach, consider how a particular disease is identified in its terms. When explaining diabetes, for example, OGMS refers quite appropriately to an "elevated level of glucose in the blood". However, it provides an insufficient account of why the explanation of diabetes needs to mention "elevated level of glucose". What role does this elevated level play in diabetes itself? Why must "elevated level of glucose in the blood" be mentioned for diabetes but nothing else? It must be something specific to the disease of interest; that is, each realization of the disease must involve an entity of this sort. For OGMS, what plays the role is the disposition and the disorder (a certain disordered body part) in which this disposition inheres. We believe that the reference to elevated level of glucose points to the need for a further type of entity, which is included in our disease model.

We know there is a difficulty in defining such a type because it is not always definite for each disease, since it varies from one patient to another. Hence OGMS' use of disposition, which is a mere potentiality. In the case of latent diabetes, for example, there is no elevated level of glucose in the blood of the patient, though there is a disposition thereto. For latent diabetes, accordingly, we follow OGMS in recognizing the need for something other than just "elevated level of glucose in the blood". But we think that there is still something more that is required – something that is essential for each particular disease. In the case of diabetes, for example, this would be the *deficiency of insulin*, since this must have happened for all patients who suffer from diabetes. To tackle this issue, we draw on OGMS' notion of homeostasis and introduce the term 'disturbance of homeostasis' to explain what we see as the essential core of each disease. Disturbance of homeostasis can be caused through the concretization of a disposition, or it can be caused through some outside agency, for example through injury.

We agree with OGMS that a disease is a dependent continuant, and its definition is expected to address the following three conditions: (1) the existence of its pre-clinical manifestation, (2) the fact that it can cause

another disease, and (3) variation in the disease course from patient to patient [1]. We try to find another definition of disease that satisfies these conditions.

3 What is a Disease?

Before going into discussion, we present some definitions of terms used in this paper. See [2] for details of event and process.

- (1) **A enacts B** =def A is a continuant and B is an external process of A who/which participates in it as a whole in which it is maximal among participants who/which play the same role in the process. Examples: when you walk, you (not your legs) enact your walking, the motion of your legs is the internal process of your walking, something which you cannot enact.
- (2) **Event** =def a non-dissective unitary entity in the temporal space. Examples include a conference, an arrival, etc. It must be dealt with as a whole in any case.
- (3) **Process** =def a dissective non-unitary entity in the temporal space like walking, singing, etc. An event is constituted of processes, unless it is instantaneous.
- (4) **External process of A** =def a process enacted by a continuant A.
- (5) **Internal process of A** =def a process enacted by a part of A. Examples: In a walking process of A, leg motion is an internal process of A whose external process is the walking.
- (6) **Causal chain** =def a chain of entities linked by causality. There can be a causal chain of disorder, causal chain of processes, causal chain of events, etc.¹

3.1 Basic Strategy

We understand a typical disease as a dependent continuant satisfies the following: After it begins to exist, it enacts extending, branching, and disappearing processes before it disappears. Thanks to these processes, a disease can be identified as a continuant that is an enactor of those processes. Such an entity (a disease) can

¹ The topic of causality is here outside our scope.

change according to its phase while keeping its identity. OGMS defines such an entity precisely as a “disposition”. Intuitively, however, it could be something related directly to a manifestation process of the disease rather than a disposition itself. At the same time, a disease should not be a process (occurrent) but a continuant. This is why defining a disease is difficult. Although the introduction of the notion of disposition is one way to solve this problem, for the reasons advanced above, disposition is a bit too far from what its manifestation process implies/suggests.

3.2 Definition

We can now define a disease as follows:

Definition 1: Disease

A disease is a dependent continuant constituted of one or more causal chains of clinical disorders appearing in a human body and initiated by at least one disorder.

Then, what is a causal chain of disorders? Although it looks like a process, it is a dependent continuant. Some people might see that a causal chain of disorders is similar to a fall of water, river flow, fire of a forest, etc. We will show how a disease is a dependent continuant rather than a process in the next section. The following is an informal account of our view.

There are two kinds of processes:

- (1) Cumulative continuous process²: a process that proceeds without completing the current process at every instant in time.
- (2) Non-cumulative process: a process that proceeds by completing the current process at every instant in time.

Most processes, such as walking, eating, talking, etc., belong to type 2. What type 1 includes are falls of water, river flows, fires of a forest, etc. The key issue here is that those cumulative continuous processes are associated with continuants, called a waterfall, a river, a forest fire, etc. in these examples. This will be briefly discussed in section 4 based on the new theory of objects, processes, and events published in [2].

A causal chain is composed of one or more

pairs of entities/events such as a causal event and an effect event, in which the latter has been caused by the former. The effect becomes another cause that causes another effect in the case of multiple-pair chains. What makes clinical causal chains special is that causal entities are usually still active when the effect entity has been caused. Therefore, the two entities overlap in temporal space. This shows that clinical causal chains belong to the type 1 process. In the case where the entities are continuants, by “an entity is active” we mean: it keeps its state as it is, so in the case of a disorder, the disordered organism still is the same disorder.

Let us examine how well a flowing river matches a causal chain of a disease. The river itself enacts branching, changing its shape, extension, diminishing, etc. In ancient times, when the river was initiated as a certain amount of water flowing, say, as a result the overflow from a lake or as a result of a heavy rainstorm, then the flow of the river is minimal. The overflow from a lake would correspond to an etiological disorder in a clinical causal chain. When the initial flow grows, the body of flowing water extends in length and is recognized as a river. After it has been born as a river (as a disease), then it extends further to another lake or to the sea. While extending, it branches (the branching perhaps causing the appearance of another disorder). Finally, it may dry up because of climate change (cure). Thus, the life of a river corresponds well to the life of a disease. Thus – in concordance with OGMS – both a river and a disease are continuants, though a river is an *independent* continuant but a disease (causal chain) is a *dependent* continuant which depends on a bearer, that is, a human being.

3.3 Discussion

On granularity: We do not specify any particular granularity of disorder and causal chains because we believe this should be flexibly determined according to the necessity of the description of each disease. Concerning the original cause, however, we have a policy that we should trace the causal chain back to the cell-level rather than to the genome-level. As far as we define diseases in general, granularity is not an issue, though it matters when we define a particular disease in the ontology.

We neither impose any specific time resolution on the causal processes so that we can

² The term *cumulative continuous process* was suggested by Barry Smith.

if needed include rapid processes such as fractures in our account. After receiving a strong external pressure, a bone undergoes a very quick destruction process resulting in fracture. The causal process can be captured by much finer time resolution than those involved ordinary pathological processes captured at the clinical level. Fracture can be dealt with by the disease model discussed in section 5.

On the distinction between a disease and a disorder: The distinction is shared with OGMS. Where OGMS defines disease (in brief) as a disposition, thus as a certain type of dependent continuant that is realized through pathological processes, we define a disease as a dependent continuant that enacts processes over pathological processes as causal chains of disorders towards a disorder(s). Disease course in OGMS is close to our definition; there, too, the disease course is a process.

4 Waterfalls and Rivers are Continuants

In order to support the above informal observation, we need to find a convincing ontological account of processes and objects. Due to space limitations, we here provide certain relevant passages from [2]. Then, we apply the discussion to support the definition of a disease as a dependent entity of a new type, different from a disposition and from a process.

Any change must be a change of something. This is already an argument against a ‘pure process’ view of reality, since we cannot conceive of processes without their material support. One might ask: what is a person over and above the sum of its internal processes? But what makes this sum worthy of consideration at all is that they constitute some kind of unity; the unity comes from the fact that there are other processes, its external processes which it enacts. Thus these questions make the mistake of focusing only on the internal processes of a person, whereas the external processes play an essential role in determining the identity of the object. Hence, rather than trying to characterize an object in terms of its internal processes (e.g., by identifying the object as the sum of those processes), we would rather say that an object is a unity which is what enacts its external processes. We could indeed say that the object is the interface between its internal and external

processes: it is a point of stability in the world in virtue of which certain processes are characterized as internal and others as external. The issue of external vs. internal processes summarizes as “The water falls, but the waterfall doesn’t fall”. That is, what a waterfall is *doing* is not *falling the water* but migrating upstream as it carves its way into the rock.

Similarly, what a river enacts is not the water flow but change of the shape of its course. This is why we can consider a river as an object that has water flow as its internal process. Similarly, a causal chain as a flow of causality (propagation of causality) is an internal process of the causal chain which is a continuant that enacts branching, extension, and diminishing processes as its external processes. Although any disease has dynamic flows of the propagation of causality as its internal processes, it is the enactor of its cumulative continuous processes such as branching and extending its causal chain of disorders as its external processes.

5 A Model of Diseases

5.1 Core Causal Chain of a Disease

On the basis of the ontological definition of diseases, we build a computational model of diseases to make it easier to define particular diseases. In the following discussion, we divide diseases into two: (1) those whose etiological and pathological processes are well-understood and (2) other diseases, and we discuss them in turn.

Diseases of type 1 are identified by their inherent etiological/pathological process(es). Diseases of type 2 include so-called syndromes and are typically represented in terms of *criteria for diagnosis*. In this section, we deal with type 1 diseases first. Let us confirm that every disease of type 1 should have a clue for identifying it. That is to say, we should be able to find something like its so-called *main pathological/etiological condition(s)* that theoretically characterizes the disease to identify it. As stated above, this is what OGMS needs to include.

We know that diseases of type 2 necessarily employ *criteria for diagnosis* to identify them because of the lack of knowledge about their etiological/pathological processes. However, this does not mean it is excluded from our disease model as is discussed below, which we share

with OGMS.

We also need a formulation for organizing diseases in an *is-a* hierarchy in a disease model. According to our definition of a disease, this would consist of a causal chain(s) which consists of nodes and links, and hence a disease is represented as a Directed Graph. We can introduce an *is-a* relation between diseases using an inclusion relationship between causal chains as follows:

Definition 2: *Is-a* Relation between Diseases

Disease A is a superclass of disease B if all of the causal chains at the class level of disease A are included in those of disease B. The inclusion of nodes (disorders) is judged by taking an *is-a* relation between the nodes into account, as well as sameness of the nodes.

Definition 3: Core Causal Chain of a Disease

The causal chain of a disease included in the chains of all its subclass diseases is called the core causal chain of the disease. An example of the core causal chain in the case of (non-latent) diabetes is:

deficiency of insulin → elevated level of glucose in the blood.

Definition 3 helps us systematically to capture the necessary and sufficient conditions of

a particular disease, which roughly corresponds to the so-called “main pathological/ etiological conditions”. Fig. 1 shows the main types of diabetes constituted by corresponding types of causal chains. The most generic type in this example is (*non-latent*) *diabetes*, which is constituted by the chain:

deficiency of insulin → elevated level of glucose in the blood

The next lower subclasses include *type-I diabetes*, which is constituted by:

destruction of pancreatic beta cells → lack of insulin I in the blood

→ deficiency of insulin → elevated level of glucose in the blood

and *steroid diabetes*, which is constituted by:

long term steroid treatment → ... → deficiency of insulin
→ elevated level of glucose in the blood

If a doctor wanted to have a hierarchy representing diabetes-caused blindness, then it would be:

deficiency of insulin → elevated level of glucose in the blood → ... → loss of sight

Due to space limitations, we omit here the discussion about the cases of resistant peripheral receptors which are also covered by our model.

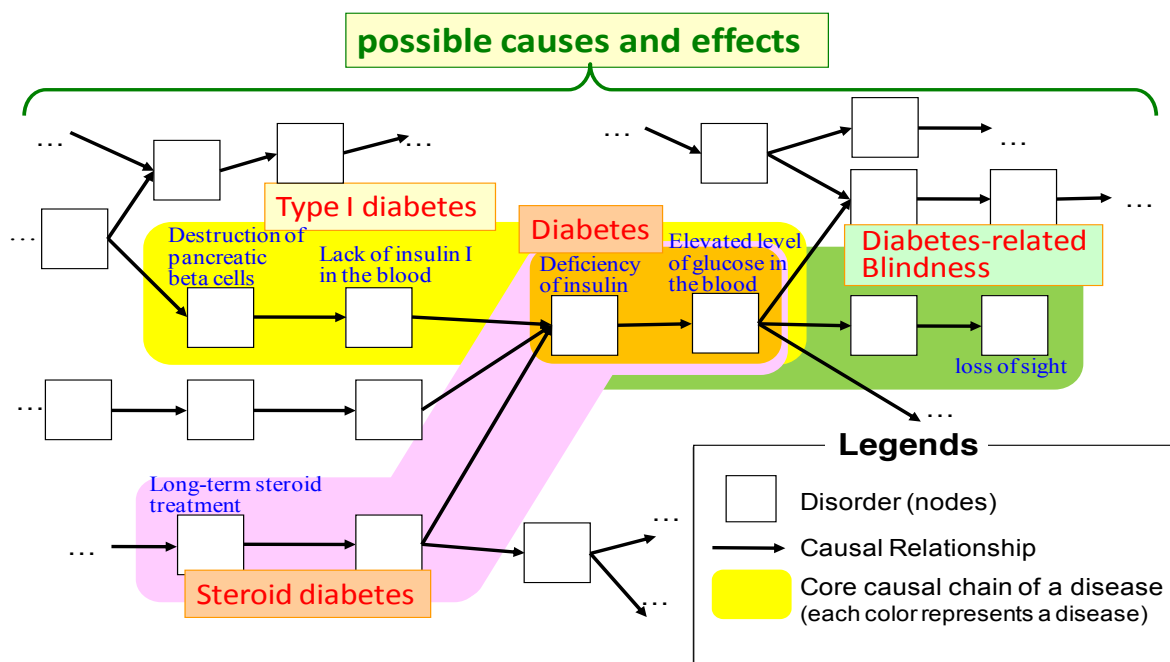


Figure 1. Types of diabetes constituted of causal chains.

Although we explain the disease model using Type 1 diseases as example, it is applicable also to Type 2 diseases thanks to the flexibility of granularity and degree of being “well-understood”. These two kinds of flexibility can be exploited according to each disease under consideration. In the case of diseases of Type 2, we could employ an “unknown” causal node linking to just a few of those symptoms that are typically observed in the case of the syndrome under consideration. Note that this model can capture a seemingly isolated symptom by combining it with an unknown cause to form a causal network. It also captures diseases with multiple causal chains.

One might suspect that this model cannot cover a phenomenon such as obesity due to the too large variety of associated causal chains, so that the classification according to Definition 2 above does not make sense. However, our model does cover obesity successfully, since it accepts multiple causal chains. Because those causal chains are not essential to obesity, unlike diabetes, they are not included in the core causal chain. Hence we do not have to classify obesity according to those causal chains. Instead, our ontology tool, HOZO [5], used for implementing the medical ontology, has a function to dynamically generate *is-a* hierarchies of diseases according to a perspective given by users [4]. Although it has some limitations, this function allows us to leave diseases in a rather flat structure if appropriate, and users classify them afterwards using the function.

Our model also can distinguish, for example, between diabetes with blindness and diabetes-driven blindness by specifying the core causal chain that is focused upon. In summary, the disease model yielded by the definition of disease proposed in section 3.2 above (Definition 1) covers quite a wide range of diseases. In fact, we have built models of 6051 diseases from 12 different divisions in our ontology, which shows the expressive power of our disease model.

5.2 Imbalance

Now, we specify the disease model discussed above by restricting diseases to deal with those of Type 1. We can introduce a mechanism to effectively model diseases of this type.

In OGMS, it seems to us that a disposition to diabetes inheres in <deficiency of insulin>. As we can easily see, <deficiency of insulin> is a physical state which, in principle, can be detected. Then, the issue is how to capture such an entity in a computational model? We have come up with the idea of *disturbance of homeostasis*. By homeostasis, we mean the same as is described in OGMS [1, p. 117]. For each parameter participating in homeostasis, there must be the notion of balance and regulation functions.

We can understand the notion of balance by introducing a performable operation (**supply**) and a required operation (**demand**). In the case of diabetes, the former is the performed amount of the insulin operation and the latter is the required amount of the insulin operation. In a normal case, the difference between the two amounts is within a certain range, that is to say, “balanced”. In an abnormal case, on the other hand, an imbalance (deficiency of insulin) occurs, which can be a disposition to the initiation of the pathological process of diabetes. On the basis of the above discussion, we define a concretized disposition as follows:

Definition 4: Clinical Imbalance

Clinical imbalance is a local phenomenon of homeostasis in the human body and is defined as a state where the difference between the amounts of supply and demand is out of the range specified for the parameter under consideration.

Precisely speaking, this balance model is represented by four nodes (see Fig. 2): balance between supply and demand, performed amount of the operation (supply), possible maximal amount of the operation, and required amount of the operation (demand). Supply can change adaptively in response to changes in the amount of demand, but only up to the maximal amount. If demand exceeds the maximal amount, a clinical imbalance occurs.

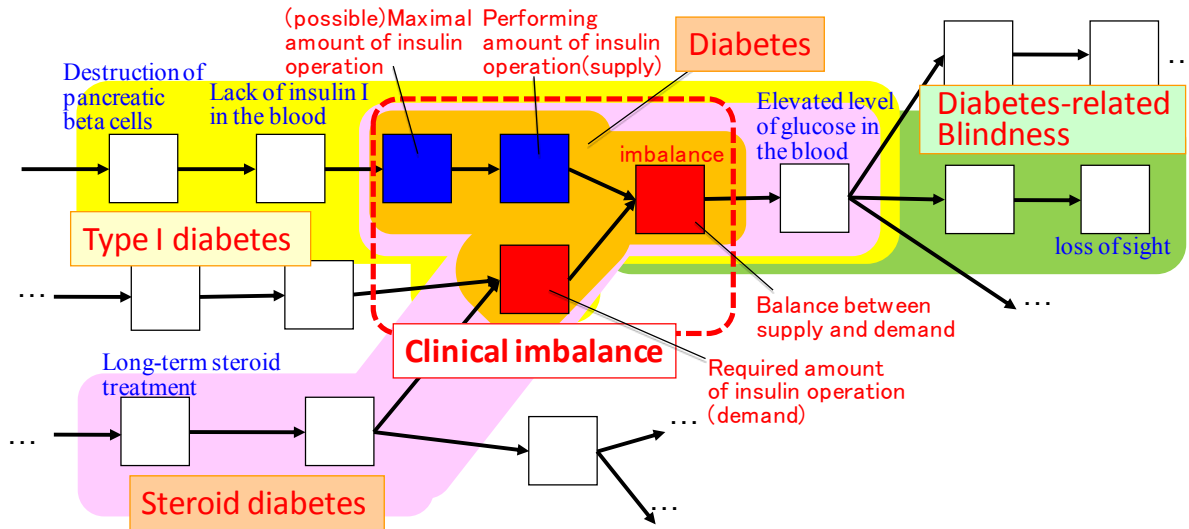


Figure 2. A representation of the clinical imbalance model

This discrimination of possible causes is critical to the proper understanding of diseases. To exploit the notion of clinical imbalance, we need some quantized generic values: **small**, **medium**, **large**, and **very large**³. By **medium**, we mean the quantity that a patient needs in everyday situations. By **small**, we mean the quantity that a patient needs in a very calm or inactive situation. By **large**, we mean the quantity that a patient needs in a stressful or active situation but that can be coped with by a normal regulation function. By **very large**, we mean the quantity that cannot be coped with by a normal regulation function. The above four nodes, except *balance*, take these four values. Due to space limitation, we have to omit an explanation of how to use the notion of clinical imbalance. Instead, we discuss the characteristics of our disease model with clinical imbalance as a factor, employing the mentioned four qualitative values.

- (1) The model can correctly capture diseases such as latent diabetes where, in OGMS terms, the relevant disposition is *not realized* at the level of clinical manifestations. In the case of latent diabetes, although maximal amount of insulin supply is **medium**, that is, smaller than those of healthy persons, the supply of insulin operation can cope only with a demand less than **large**. Therefore, no

imbalance occurs while the patient is going about her normal daily activities. However, the fact that the maximal amount is smaller than **large** shows that the patient is said to suffer from latent diabetes. If the demand for insulin is increased for some reason and becomes greater than the maximal amount of insulin supply, then imbalance occurs, and the diabetes is no longer latent.

- (2) We evaluated the expressive power of the model by representing diabetes, ischemic heart diseases, infectious diseases, and osteoporosis and found it worked satisfactorily. For example, fracture caused by osteoporosis is modeled using **medium** value for demand to resist the normal pressure given as an external cause and **small** value for supply to resist external pressure. While bones of normal people can stand such external pressure, patients that suffer from osteoporosis cannot, which the imbalance model clearly explains.
- (3) These four qualitative values work well to represent each particular disease whose instances share the same threshold values to quantize real values, though it does not make sense to compare them across different diseases. Because our goal is defining each particular disease rather than diagnosis, we do not need concrete thresholds or ranges of their values.

³ Although it is explained in terms of the *demand* side in the following, it also applies to the *supply* side in a similar way.

- (4) The parameter(s) chosen in the model should be dependent on the particular disease under consideration but causes no problem, since it is what the medical experts think essential for capturing the disease.

6 Concluding Remarks

We have discussed a definition of a disease friendly to domain experts based on a new ontological theory of objects and processes. We conducted a small informal evaluation by asking seven medical doctors with different expertise who are totally unfamiliar to ontology and do not know the authors which definition they like better among the two definitions and learned all selected our definition, which suggests our definition is friendly to them more than that in OGMS.

The definition enables us to understand a disease as a dependent continuant constituted of a clinical causal chain(s). We also discussed a model of a disease that allows the definition to be implemented. In the model, we defined a core causal chain of a disease with the idea of clinical imbalance which can be considered as the concretization of disposition to a disease. With this approach, we believe that we have improved OGMS. We have been developing a medical ontology for the last four years on the basis of our definition and model of diseases [4]. As of April 12, 2011, a total of 6051 diseases have been defined in the medical ontology by 12 clinicians, and these definitions are currently being refined. Currently, we have no concrete connection with activities conducted outside

Japan, but we are open for collaboration with ventures as DO [6], OBO[7] as well as with OGMS.

Acknowledgement

A part of this research is supported by the Japan Society for the Promotion of Science (JSPS) through its Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program). The authors are grateful to all the reviewers, who gave us valuable and constructive comments.

References

1. Richard H. Scheuermann, Werner Ceusters, and Barry Smith, Toward an Ontological Treatment of Disease and Diagnosis, Proceedings of the 2009 AMIA Summit on Translational Bioinformatics, San Francisco, CA, 2009. pp. 116-120 (2009)
2. Antony Galton and Riichiro Mizoguchi, The water falls but the waterfall does not fall: New perspectives on objects, processes and events, *Applied Ontology* 4(2), pp. 71-107 (2009)
3. BFO: <http://www.ifomis.org/bfo/>
4. Riichiro Mizoguchi, Hiroko Kou, Jun Zhou, Kouji Kozaki, Ken Imai and Kazuhiko Ohe, An Advanced Clinical Ontology, Proc. of International Conference on Biomedical Ontology (ICBO), Buffalo, NY, June 24-26, 2009, pp. 119-122 (2009)
5. HOZO: <http://www.hozo.jp/>
6. DO, http://do-wiki.nubic.northwestern.edu/index.php/Main_Page/
7. OBO: <http://www.obofoundry.org>

Dispositions and Processes in the Emotion Ontology

Janna Hastings^{1,2}, Werner Ceusters³, Barry Smith⁴, Kevin Mulligan¹

¹Department of Philosophy and Swiss Centre for Affective Sciences, University of Geneva, Geneva, Switzerland

²Chemoinformatics and Metabolism, European Bioinformatics Institute, Hinxton, UK

³Department of Psychiatry and Ontology Research Group, New York State Center of Excellence in Bioinformatics & Life Sciences, University at Buffalo, NY, USA

⁴Department of Philosophy and National Center for Ontological Research, New York State Center of Excellence in Bioinformatics & Life Sciences, University at Buffalo, NY, USA

Abstract. Affective science conducts interdisciplinary research into the emotions and other affective phenomena. Currently, such research is hampered by the lack of common definitions of terms used to describe, categorise and report both individual emotional experiences and the results of scientific investigations of such experiences. High quality ontologies provide formal definitions for types of entities in reality and for the relationships between such entities, definitions which can be used to disambiguate and unify data across different disciplines. Heretofore, there has been little effort directed towards such formal representation for affective phenomena, in part because of widespread debates within the affective science community on matters of definition and categorization. We describe our efforts towards developing an Emotion Ontology (EMO) to serve the affective science community. We here focus on conformity to the BFO upper ontology and disambiguation of polysemous terminology. The full ontology is available for download from: <https://emotion-ontology.googlecode.com/svn/trunk/ontology/EMO.owl> under the Creative Commons CC BY 3.0 Attribution license.

Introduction

High quality ontologies in the biomedical sciences enhance the potential for integration of the exploding quantities of experimental and clinical data that have become available on-line. When appropriately designed, ontologies allow annotations of data to be unified through disambiguation of the terms employed in a way that allows complex statistical and other analyses to be performed which lead to the computational discovery of novel insights [11].

Affective science is the study of emotions and of affective phenomena such as moods, affects and bodily feelings. It combines the perspectives of many disciplines, such as neuroscience, psychology and philosophy [2]. Emotions have a deep and profound influence on all aspects of human functioning, and altered or dysfunctional emotional responses are implicated in both the etiology and the symptomology of many pathological conditions. Depression, for example, which is characterised by abnormally low affect and generally flattened emotional reactions, is one of the fastest-growing public health

problems in many countries, corresponding to massive growth in sales of pharmaceuticals (and other substances) which target human affect [9].

Research in affective science faces the need to integrate results obtained on the basis of subjective reports and those obtained by neuroscientific or other methodologies, and to compare results across disciplines. It is therefore essential to have a shared, disambiguated and clear reference terminology for the domain [4, 13]. To address this requirement, we are developing an Emotion Ontology (EMO). Due to space constraints, we focus in this paper on addressing certain ambiguities in the language we use to talk about emotions and on providing definitions for one variety of affective phenomenon, namely occurrent emotions.

1 Background

1.1 Ambiguity in Emotion Language

Emotion terminology in English displays the following two-tiered structure [20]. First of all, there are a handful of fairly high-level terms: *affect*, *feeling*, *emotion*, *mood*, *passion*, *sentiment*.

Secondly, there are a large number of much more concrete terms for particular emotion types such as *anger*, *astonishment*, *awe*, *bliss*, *despair*, *disgust*, *embarrassment*, *fear*, *happiness*, *hate*, *joy*, *love*, *pride*, *regret*, *resentment*, *satisfaction*, *scorn*, *shame*, *sympathy* and *terror*. A similar structure is to be found in many other natural languages [3].

This emotion terminology may be used to describe phenomena of different sorts, as we can see by considering examples for the emotion *anger*.

A statement that John is angry with Mary, in the absence of any further contextual clues, can be interpreted to mean (Scenario 1) that John is experiencing a feeling of anger and displaying anger behaviour which has Mary as its object or target, that is: he is speaking in a raised voice to or about her, clenching his fists, breathing heavily and so on. The outward behaviour which expresses the anger may, to a certain extent, be suppressed, but internally, the anger is strong. In this scenario, the anger that John is experiencing and displaying is an occurrent entity (a process).

On the other hand, the statement that John is angry with Mary may be interpreted to mean (Scenario 2) that John is likely to react angrily, for example due to some past slight, given the right triggers, such as if Mary comes into the room, or her name is mentioned in conversation, causing his anger at Mary to flare up again, *even though* he was quite happy beforehand, while thoughts of Mary were far from his mind. Here, the anger that John has for Mary is dispositional. In such a case John is angry even when he feels no anger [10, 3]. Note that this is a different disposition from the one for John to become angry *in general*: this can be seen by considering that John would not normally become angry with someone just because they walk into the room. Rather, his anger when Mary does so is caused by a specialization of his general disposition to become angry, just as driving this car or that truck are specializations of driving in general [6]. This distinction is of relevance in research involving self-reports, since subjects may self-report being angry in the dispositional sense when the experimenters are attempting to analyse characteristic brain states related to anger processes. Another complication is that John may or may not be *aware* of his anger.

Finally, emotion language can be used to describe stable or enduring personality traits (Scenario 3), such as in the statement that John is an angry person. This distinction may again be important for annotation of neuroscientific research data, because researchers may well wish to differentiate persons who generally have low anger thresholds from those who have more calm temperaments.

We will use the following terminology for the above scenarios:

- Scenario 1: *Emotion occurrent*. An emotion occurrent is a processual emotion which in which a person participates over a specific time period. A person undergoes or is the subject of the emotion; he – we might say – *emotes*. This terminology leaves open what the person feels or is aware of.
- Scenario 2: *Emotion disposition*. An emotion disposition is a disposition to undergo emotion occurrences if the right circumstances obtain.
- Scenario 3: *Emotional personality trait (predisposition)*: An emotional personality trait is a stable enduring characteristic of a person which involves a predisposition (i.e. a disposition which gives rise to an increased risk) to undergo emotions of a particular sort, both occurrences and dispositions.

Both the emotion disposition and the emotion personality trait are dispositions which are *realized* in emotion occurrences.

1.2 Basic Formal Ontology

EMO is being developed beneath the Basic Formal Ontology (BFO) [17, 5]. BFO distinguishes between *occurrences* and *continuants*. *Occurrences* are those entities that unfold over time and have temporal parts; *continuants* are those entities that endure through time and are wholly present at all times that they exist. For example, John is a *continuant*, but his angry behaviour is an *occurrence*.

The *continuant* branch is then further sub-divided between those entities which exist independently, such as John, and those which are ontologically dependent on some other entity for their existence, such as John's personality (which depends on John). Dependent entities may be *qualities*, such as colour, or they may be

realizable entities, such as dispositions, which are entities that inhere in other entities and which have the nature of propensities or potentials by virtue of which occurrents of certain sorts will be realised if the underlying bearer entity comes into the right *circumstances*. In BFO, dispositions are the most general class of such propensities, with subtypes for tendencies and “surefire” dispositions. An example of a disposition is John’s disposition to become angry, for instance when Mary’s name is mentioned.

1.3 Emotions as Componential Processes Caused by Appraisals

The question ‘What is an emotion?’ has been widely debated in both philosophy and science, with different accounts focusing on the centrality and essentiality of different aspects of emotional experiences [4, 8, 10, 13].

Appraisal theories are a modern variant of cognitive theories of emotion [14]. A central claim of cognitive theories is that representations are constitutive of emotions – that is, that emotions are identified by, and caused by, cognitive representations [8]. The main alternative to cognitive theories are feeling theories, which embrace a type of view which goes back to William James and which identifies episodic emotions with bodily feelings and sensations or with the awareness of these [10]. Although a great deal of ink has been spilt in discussions of theories on the continuum between cognitive and feeling theories, there is extensive agreement about the nature of many of the phenomena which figure in these theories – cognitive theories admit that emotions are often accompanied by subjective feelings, and feeling theories admit that emotions are often accompanied by cognitive representations – and disagreements revolve mainly around the underlying mechanisms and causal pathways

[4]. It is this common ground which will serve as a starting point for the development of EMO, with future work elaborating the areas of disagreement. A representative definition of ‘emotion’ is [12, 13]:

An episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism.

The ‘episode of interrelated changes’ here corresponds to what we have called an ‘occurrent emotion’ in the above discussion. Scherer goes on to distinguish five different components essential to emotion occurrents, related to the five organismic subsystems as listed in Table 1 and illustrated in Figure 1 [12, 13].

2 Foundational Entities in the Emotion Ontology

EMO describes and places emotion occurrents together with their five component parts. To align with BFO, we will draw on terminology already defined in the Ontology of Mental Disease (OMD) [1]. OMD starts out from the view that all mental disease rests on some physical disorder in the patient, such as a portion of the brain that is affected by some chemical imbalance or that has become damaged due to injury, and then aims to describe what it is for something to be a mental disease. Relevant entities from OMD which will be used in EMO are illustrated in Figure 2.

We can now begin to place the emotion entities beneath this framework via subtyping. We will divide the entities between those that are processual, those that are representations, and those that are dispositions.

Emotion component	Function	Organismic subsystem	Major substrata
Appraisal	Evaluation of objects and events	Information processing	CNS
Neurophysiological component	Bodily symptoms, system regulation	Support	CNS, ANS, NES
Action tendencies	Preparation for action	Executive	CNS
Motor expression behaviour	Communication of reaction and intention	Action	SNS
Subjective feeling	Monitoring of internal state	Monitor	CNS

Table 1. Component parts of emotion occurrents.

CNS – Central Nervous System. ANS – Autonomic Nervous System. SES – Somatic Nervous System. NES – Neuro-Endocrine System.

The **appraisal** is the evaluation of a stimulus event as *relevant* to the organism (CNS)



Subjective feeling involves the subjective experience of the emotion (CNS)

Behaviour involves the characteristic facial and vocal expression changes for the emotion, controlled by the somatic nervous system (SNS)

Action tendencies involve the motivational aspects elicited by the appraisal (CNS)

Physiological response encompasses the *neurophysiological changes* which take place, e.g. in the central nervous system (CNS), neuro-endocrine system (NES) and autonomous nervous system (ANS)

Figure 1. Components of John's anger occurrent

2.1 Mental and Physical Processes

An emotion occurrent is a synchronized complex of mental and physical processes. More specifically:

An *emotion occurrent* is a mental process that is a synchronized complex of constituent mental and physical processes including an *appraisal process* as part, and which gives rise to an *action tendency*. At least one appraisal precedes the other components of the emotion, while it or others continue throughout the emotion occurrent and guide the process.

The constituent mental and physical processes that usually form a part of the emotion occurrent include the appraisal process, bodily changes (such as increased heart rate), experience of subjective feelings, and behavioural processes (such as altered facial expressions). Different emotions are characterised by differing patterns of behaviour and action tendencies.

An *appraisal process* is a mental process that gives rise to an *appraisal*, which will be defined below.

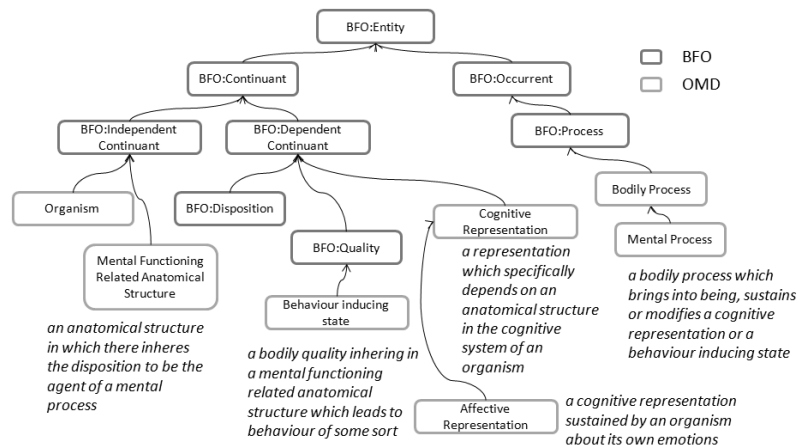
A *physiological response to emotion process* is a bodily process which encompasses all the neurophysiological changes caused by the emotion, which take place in the central nervous system (CNS), neuro-endocrine system (NES) and autonomous nervous system (ANS).

An *emotional behavioural process* is the behaviour of the organism in response to the emotion, which includes the characteristic facial expressions for particular emotion types.

Expressive behaviour is an actual part of an emotion, and actual actions seem to be part of some emotions, such as *anger* and *fight* or *flight*. But not of all:

For pride and humility are pure emotions in the soul, unattended with any desire, and not immediately exciting us to action. But love and hatred are not completed within themselves, nor rest in that emotion, which they produce, but carry the mind to something further. (Hume Treatise Book II Section VI p. 115)

Figure 2. Overview of the top level of the Ontology of Mental Disease. Arrows represent 'is a' relations.



2.2 Mental Representations

Core to appraisal theories is the centrality of the appraisal as a mental representation.

An *appraisal* is a cognitive representation which represents an evaluation of the *relevance* of some triggering object or event to the organism.

Appraisals represent value relations (good, bad, better...) which are judged as holding between the appraising organism and the triggering object or event. If Sam is afraid of the dog one object of his fear is the dog (the triggering object), the other is its dangerousness for him (a kind of 'bad' value relation). If John is angry with Mary, one object of his appraisal is Mary, the other is his judgment of Mary's behavior as unjust or insulting towards him.

Appraisals are normally taken to be judgments which use *concepts*, which implies that non-concept-forming organisms (such as small babies and dogs) cannot have emotions. Some researchers therefore assign a different type of emotion in this case, called a *proto-emotion* [8], which lacks a full appraisal component. In EMO, in line with multi-level appraisal theorists [14], we presuppose only that occurrent emotions are experiences of a *type that can be* associated with complex concepts in those organisms which can form concepts. The complexity of emotions which can be experienced by organisms increases with the conceptual complexity of the organism – we do not ascribe complex emotions such as *schadenfreude* (taking pleasure in the misfortune of others) to babies or animals.

Just as one may be aware of one's visual perceptions, memories or judgments so too one may be aware of one's anger. The subjective feeling of the emotion is also a representation, but of a different sort. Awareness of one's emotion is a type of inner perception. Just as one can be aware of one's sadness, one can be aware of seeing someone. Such awareness and inner perception are intentional acts or states, directed towards objects. It is often assumed that feeling is a awareness or inner perception, but in fact the most common locution in this area is *X feels sad/unhappy/angry*. If ordinary language is a reliable guide, therefore, feeling sad is a *way* of feeling ("sad"), not *what* one feels ("sadness"). In what follows we leave open the exact relation between these two types of feeling.

The *subjective emotional feeling* is an *affective representation*, that is, a representation that the organism has about its own affect.

An affective representation differs from a cognitive representation, since the latter can be divided between a form of judging or belief and a form of inner perception, while affective representations are a fusion of these: to be aware of one's anger is to feel anger. Furthermore, they differ in their objects: the affective representation is about the *internal state* of the organism, while the appraisal is about the relevance of the triggering object or event to the organism.

Examples of subjective emotional feelings are the characteristic angry feeling accompanying anger, or the highly painful feeling that accompanies grief.

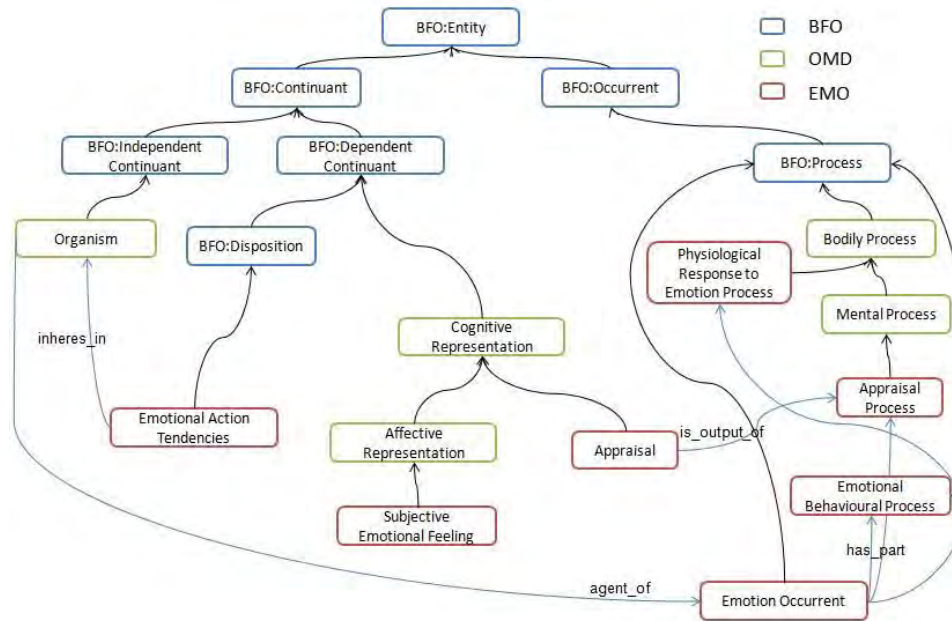


Figure 3. Processes, representations and dispositions in the Emotion Ontology. Unlabelled arrows represent ‘is a’ relations.

2.3 Dispositions

The final entity in our list of emotion components is the action tendencies which are elicited by the emotion. Action tendencies differ from behaviour in that they may not be realized; indeed, action tendencies, if they are realized, are realized as behaviour. For this reason, we say they are *dispositions*. Dispositions and tendencies are not parts of emotions since they are not occurrents, but they are parts of the definitions of many emotions.

Emotional action tendencies are dispositions to behaviour which inhere in an organism by virtue of the physical changes brought about by an emotion process.

The resulting ontology is illustrated in Figure 3.

3 Discussion

Recognising the need for clear categorical distinctions in support of research design, the accumulation of research findings, and linking affective science to the biomedical science of affective disorders, emotion researchers have long been proposing typologies and lists of emotions and affective phenomena [4, 10, 8, 13]. Thus far, a broad shared agreement on definitions for emotion terms has not been achieved, although there is agreement on many

of the relevant constituent elements [4].

Requirements in computing and artificial intelligence have led to the development of ontology-like resources for emotions. *Affective computing* aims to integrate emotional responses into computer interfaces in order to produce more realistic systems which are able to respond to the emotional communication of their users. To facilitate affective computing, Lopez *et al.* propose a slim ontology schema for describing emotions in *human-computer interfaces* [7]. Also motivated by affective computing requirements, the W3C’s emotion markup language (EML, <http://www.w3.org/TR/emotionml/>) is an XML-based standard for markup of emotions in text or databases. Another computing application which has led to developments in this domain is natural language processing, for which Valitutti and Stock developed an emotion lexicon [19], Triezenberg has developed an emotion terminology which categorises emotion types and related behaviour [18], and Yan *et al.* have developed an extensive terminology for the domain of emotions as expressed in Chinese [21]. Effectively marking up references to emotions in text, databases, and human-computer interfaces relies on an unambiguous shared understanding of what emotion terms denote. All of the ontology-like resources that have thus far been developed make use of

emotion terms assumed to be defined elsewhere. The formal and unambiguous scientific definition for terms in this domain is therefore still an open requirement, and it is to fill this gap that the power of shared community-wide ontologies is required.

Our approach, following best practices promoted by the OBO Foundry [15] and the principles of Ontological Realism [16] will be to engage each of the different sub-communities, both scientific and computational, at every stage in the development of EMO in order to address and reconcile, rather than ignore, fundamental terminological and definitional disagreements. This will allow the application of the developed ontology to multiple application scenarios both in support of scientific research and in support of intelligent computing.

4 Conclusion

We have presented the first steps in the development of an emotion ontology based on BFO and OMD, focusing on the definition of occurrent emotions and their component parts. Researchers continue to disagree on the essential nature of emotions. For example, behaviourists give central importance to expressive behaviour in defining emotions, while some cognitive theories take actual behaviour to be essential only to some emotions, such as anger. To address this, our ontology will provide definitions for the different terms used to denote *components* of emotions, which can then be used separately if needed in annotations of the relevant scientific literature.

Much future work remains in the project, first in terms of enhancing the ontology through delineating the different types of emotions and defining appropriate terms for them, and also addressing the different types of affective phenomena of other sorts, such as moods, and secondly in terms of applying the ontology in the practical annotation of scientific research data and in the development of novel applications that draw on emotion semantics.

Acknowledgements

This work was supported in part by the Swiss National Science Foundation, and by the US National Institutes of Health under Roadmap Grant 1 U 54 HG004028. We are grateful to

Damiano Costa from the University of Geneva and Colin Batchelor from the RSC in Cambridge for helpful comments.

References

1. Ceusters, W., Smith, B.: Foundations for a realist ontology of mental disease. *Journal of Biomedical Semantics* 1(1), 10 (2010)
2. Davidson, R.J., Scherer, K.R., Goldsmith, H.H.: *Handbook of Affective Sciences*, vol. Series in Affective Sciences. Oxford University Press (2003)
3. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.: The world of emotion is not two-dimensional. *Psychological Science* 18, 1050–1057 (2007)
4. Frijda, N.H., Scherer, K.R.: *Emotion definitions (psychological perspectives)*. Oxford University Press, New York (2009)
5. Grenon, P., Smith, B., Goldberg, L.: Biodynamic ontology: Applying BFO in the biomedical domain. In: *Stud. Health Technol. Inform. pp. 20–38*. IOS Press (2004)
6. Johansson, I.: Four kinds of “is a” relations: genus-subsumption, determinablesubsumption, specification and specialization. In: Johansson, I., Klein, B. (eds.) *WSPI 2006: Contributions to the Third International Workshop on Philosophy and Informatics*, Saarbrücken, May 3–4, 2006 (2006)
7. López, J., Gil, R., García, R., Cearreta, I., Garay, N.: Towards an ontology for describing emotions. In: Lytras, M., Carroll, J., Damiani, E., Tennyson, R. (eds.) *Emerging Technologies and Information Systems for the Knowledge Society*, LNCS, vol. 5288, pp. 96–104. Springer Berlin / Heidelberg (2008)
8. Lyons, W.E.: *Emotion*. Cambridge University Press (1980)
9. Patel, V., Flisher, A.J., Hetrick, S., McGorry, P.: Mental health of young people: a global public-health challenge. *The Lancet* 369(9569), 1302 – 1313 (2007)
10. Prinz, J.J.: *Gut Reactions: A Perceptual Theory of the Emotions*. Oxford University Press (2004)
11. Rubin, D.L., Shah, N.H., Noy, N.F.: Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics* 9(1), 75–90 (2008)
12. Sander, D., Grandjean, D., Scherer, K.R.: A systems approach to appraisal mechanisms in emotion. *Neural Netw.* 18(4), 317–352 (May 2005)
13. Scherer, K.R.: What are emotions? and how can they be measured? *Social Science Information* 44, 695–729 (2005)

14. Scherer, K.R., Ellsworth, P.C.: *Appraisal theories*. Oxford University Press, New York (2009)
15. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25(11), 1251–1255 (Nov 2007)
16. Smith, B., Ceusters, W.: Ontological realism as a methodology for coordinated evolution of scientific ontologies. *Applied Ontology* 5, 139–188 (2010)
17. Smith, B., Grenon, P.: The cornucopia of formal ontological relations. *Dialectica* 58, 279–296 (2004)
18. Triezenberg, K.: *The Ontology of Emotion*. Ph.D. thesis, College of Liberal Arts, Purdue University (2005)
19. Valitutti, R., Stock, O.: Developing affective lexical resources. *PsychNology Journal* pp. 61–83 (2004)
20. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39, 165–210 (2005), 10.1007/s10579-005-7880-9
21. Yan, J., Bracewell, D.B., Ren, F., Kuroiwa, S.: The creation of a Chinese Emotion Ontology based on HowNet. *Engineering Letters* 16 (2008)

An Advanced Strategy for Integration of Biological Measurement Data

Hiroshi Masuya¹, Georgios V Gkoutos², Nobuhiko Tanaka¹, Kazunori Waki¹, Yoshihiro Okuda³,
Tatsuya Kushida³, Norio Kobayashi⁴, Koji Doi⁴, Kouji Kozaki⁵, Robert Hoehndorf²,
Shigeharu Wakana¹, Tetsuro Toyoda⁴, Riichiro Mizoguchi⁵

¹RIKEN BioResource Center, Tsukuba, Japan; ²Department of Genetics, University of Cambridge, UK;

³NalaPro Technologies, Inc, Tokyo, Japan; ⁴RIKEN BASE, Yokohama Japan;

⁵Department of Knowledge Systems, ISIR, Osaka University, Ibaraki, Japan

Abstract. Aiming at integration of measurement data across various biological experiments, we investigated a methodology for expanding the Phenotypic Quality Ontology (PATO), commonly used for descriptions of biological phenotypes, based on the YAMATO top-level ontology. The mapping of ontology terms from PATO to the YAMATO framework brings several benefits, including: introduction of a classification of quality values to represent measurement scales; distinction of different contexts in which comparisons of ordinal values are made; and establishment of interoperability of quality description formalisms based on different top-level ontologies. In this study, we propose an ontological basis for integrating cross-species and cross-experimental biological measurement data.

Keywords: Interoperability, top-level ontology, quality, phenotype

1 Introduction

A phenotype is an observable and measurable quality of a biological entity. Phenotypes represent a broad range of variations in measured qualities along dimensions such as morphology, development, biochemical or physiological properties, behavior, and so on. For a better understanding of living organisms at the systemic level, it is essential to integrate phenotypic information at all levels and along all such dimensions. This requires the development of a sophisticated informatics infrastructure for the description, exchange and integration of phenotypic data. The ontological formalization of the description of phenotypic qualities is a core issue in the development of such an infrastructure.

To realize a high degree of freedom of data representation, phenotypes are often described by means of representations that employ ontology terms [1]. The Descriptive Ontology for Linguistic, Cognitive Engineering (DOLCE) [2], in contrast, uses the <Entity, Attribute, Quality value> (EAV) formalism, as in <John, height, 180 cm>. The Generalized Architecture for Languages, Encyclopedias and Nomencla-

tures (GALEN) [3] employs the <Entity, Property, Quality value> (EPV) formalism, as in <John, height, tall>. The EAV triple is used also in [4], which employs Minsky's frame-based knowledge representation, the *de facto* standard in the field of artificial intelligence studies. The EAV triple distinguishes between the quality and quality value [2].

Within the Open Biomedical Ontology (OBO) community, the Phenotype Quality Ontology (PATO) provides a practical basis for vocabulary and semantics for the description of phenotype information across species [5]. PATO follows the Basic Formal Ontology (BFO) framework, [6] which recommends a variant of the <Entity, Property> formalism (EP), as for example in <John, height>, <John, above average height>, <John, 220 cm height>. This yields what are called "entity quality" (EQ) annotations of experimental parameters and parameter values; the EQ is an equivalent of the EP formalism. In contrast to the EAV and EPV approaches, this implies a single hierarchy of qualities, rather than a double hierarchy of both qualities and values.

PATO has contributed greatly to the development of a practical basis for the

qualitative and quantitative description of biological phenotypes. However, a number of problems still remain to be resolved, including the problem of classification of quality values, of measurements made in different contexts, and of variations in the descriptions of quality-related information created by separate research communities.

There are various efforts to improve ontologies of measurement and of qualities [7-9]. In this study, we attempted to expand the PATO ontology to ensure a more advanced framework within the YAMATA (Yet Another More Advanced Top-Level Ontology) framework [10]. Our goal is to integrate quality descriptions deriving from experimental studies in biomedicine through the development of a reference ontology named “PATO2YAMATO”.

2 Practical Requirements of an Ontological Basis for Describing Biological Measurements

2.1 Fundamental Classification of Quality Values on the Basis of Measurement Scales

In the field of experimental biology, the results of measurements are described in terms of a variety of systems of “values”. One of the most common classifications of values is “levels” or “scales” of measurement developed by Stanley S. Stevens [11]. This describes four different types of scales, namely “nominal (categorical)”, “ordinal”, “interval,” and “ratio”. This classification takes as its starting point the mathematical operation that is applied in analyses of measurement results. Therefore, such a classification is quite beneficial for data integration in the field of experimental biology. BFO provides a single node quality, and quality ontologies such as PATO are created by downward population from this single node (hence PATO is referred to as adopting a single hierarchy approach). DOLCE, in contrast, provides a bi-hierarchical system of classification for quality-related concepts involving both *quality-space* and *quale*, providing a classification of both the type of quality involved (in terms of *quality*

space) and of the value or *quale*. Integration of qualitative and quantitative descriptions is realized by a combination of quality-space and *quale* in a single knowledge framework. As mentioned elsewhere [10], there seems to be some room for improvement in DOLCE in regard to the current running together or unclarity in its treatment of value, quality of real entity, and data about this quality.

2.3 Modeling of Context Dependencies of Ordinal Scale Values

In general, an ordinal value is used within a certain sort of phenotypic quality description when making comparison for example of *normal* versus *abnormal*, for instance in regard to some developmental or evolutionary change. The problem is that different contexts, here, can lead to different bases for comparison, illustrated by the fact that from *large ant*(x) and *small elephant*(y) we could never infer that x is larger than y. “Abnormally large” and “abnormally small”, which are equivalent to “increased size” (PATO:0000586) and “decreased size” (PATO:0000587), refer to groupings of values based on deviation from the normal in specific biological populations such as species or strains. This yields a completely different view of classification from that which is obtained from a simple grouping into “small” and “large” (Fig 1).

On the other hand, when comparisons are based on deviations within distinct species are closely related to each other. For example, “abnormally largeness” of homologous skeletal elements in humans and in mice are often caused by mutations of homologous genes. This indicates that it might make sense to consider “abnormally large in ant” and “abnormally large in elephant” as subclasses of a single class “abnormally large”. There are analogous interrelationships between classifications of biological species and of experimental population, and so on. To establish advanced integration of biological measurements, it is essential to provide an explicit and consistent way to document such interrelationships.

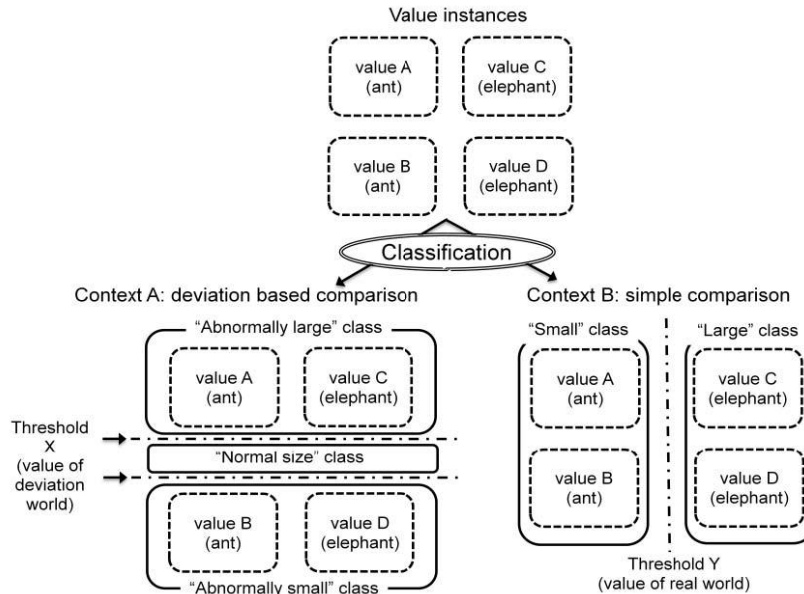


Figure 1. The problem of “large ant and small elephant”.

Rounded squares with broken lines and solid lines represent instances and classes of quality values respectively.

2.3 Modeling of a Datum as an Informational Entity

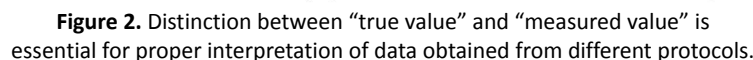
Empirical measurements are always approximations that do not accurately represent what is being measured. In other words, the “true value”, a quality (value) that is independent of any act of measurement, is clearly distinct from the “measured value” (or “datum”), a representation or description of the quality obtained through such an act. As pointed out already in [10], it is important for the integration of biological measurements to define a “datum describing a quality” as an informational entity that is related with the quality itself by some relation of aboutness.

While the acknowledgement of two entities – a datum to describe the quality in addition to the quality itself – seems to imply redundancy, this move is essential for any strategy leading to the integration of biological measurement data. To see why, consider the following example of a case of measurement of body weight (W_n) of a mouse performed using different procedures (A and B) along a time course (T_n). If all measured values are regarded as true values, the changes of the mouse’s body weight would be recorded as:

$$w1(a, t1) \quad w2(b, t2) \quad w3(a, t3) \quad w4(b, t4)$$

and so on (see Fig. 2A). In the case in hand, however, this will likely lead to a wrong interpretation of the changes in the body weight. A more standard interpretation would consider the two series of results, $w_i(a, t_i)$ and $w_j(b, t_j)$ independently, and then form an estimate of the true values of the mouse’s weight at successive times, as in (Fig. 2B).

Within the bio-ontology community, phenotype annotations are usually recorded following the EQ or EAV formalisms. These formalisms are convenient to represent qualities in the biological measurements. In order to deal with symbolically formalized information items, an ontology of representation have been proposed in [12]. The Information Artifact Ontology (IAO) [13] also defines measurement data and descriptions as information entities to be related to the real-world in biomedical investigation. The sophisticated modeling and classification of “formalisms” to be used in the informational items and the facilitating the interoperability between these formalisms are then a further issue that must be addressed for the purposes of integration of biological measurement data.



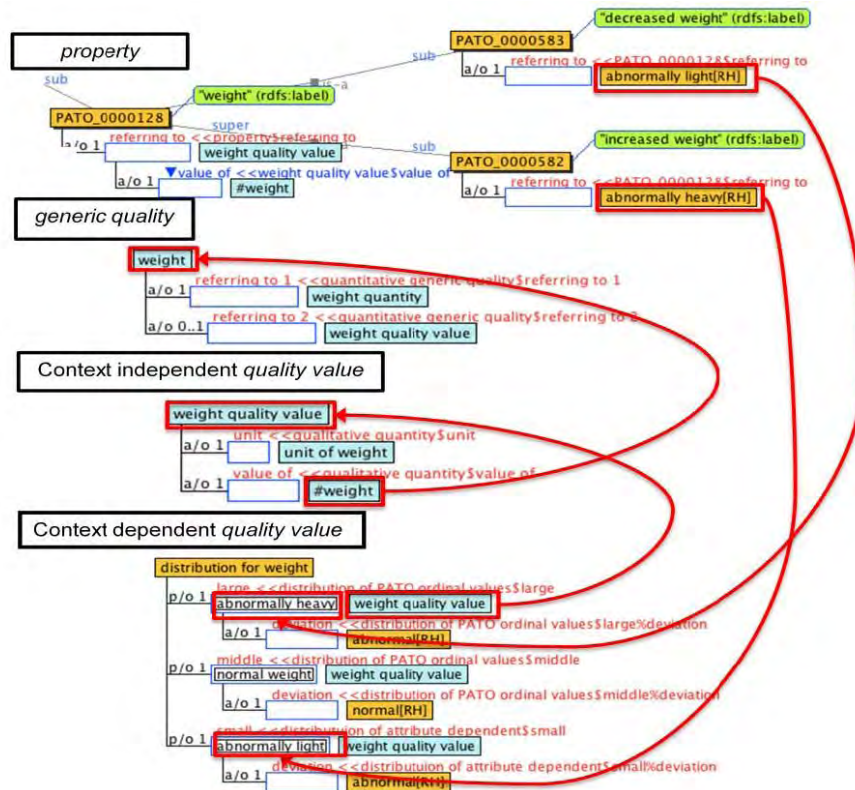


Figure 4. Example of the inter-relationships between imported original terms and newly defined terms in PATO2YAMATO. Orange and blue boxes represent terms to be incorporated into PATO2YAMATO and YAMATO respectively.

4 Summary of the Semantics of the Resultant Ontology

Figure 4 shows an example of the inter-relationships between imported original terms and newly defined terms. In the YAMATO framework, PATO:0000582 (increased weight) is defined as a *Property* (equivalent to BFO's *Quality*) that is philosophically a combination of a *Generic quality* (type of quality), *weight*, and a context-dependent *Quality value*: *Abnormally heavy*.

The context-independent value is defined as a class “*Weight quality value*”. This class is instantiated in each specific context. In Figure 4, “*Increased weight*” is defined as being instantiated in the context, “*Distribution of*

weight”. This can be rephrased in the statement that “in the **distribution of weight**, some **weight quality values** playing **large**-roles thereby becomes role holders, **abnormally heavy**”

The classification of contexts for the ordinal values is illustrated in Figure 5. The context, “*Abnormally large, small and normal*” is the three-value comparison based on the deviations of abnormally large and abnormally small from normal values. This is clearly distinguished from the simple three-value comparison (“*Simple large, middle and small*” in Fig. 5). These values depend on the context of “*Distribution of ordinal values*”, which is given by each species.

provides one of the basic procedures essential for a number of operations in the integration of biological measurement data to integrate the features of multiple top-level ontologies such as DOLCE and BFO. The Open Biomedical Ontologies (OBO) Foundry has coordinated the definition of scientific methods to be applied in ontological developments [20]. Toward forming a single, consistent, cumulatively expanding and algorithmically tractable whole, the OBO Foundry applied only BFO as the semantic framework. However, BFO itself has undergone and will continue to undergo modifications over time. Already for this reason, therefore, we believe that investigation with multiple top-level ontologies would facilitate not only YAMATO-based integration, but also the OBO foundry initiative to show the requirements and concrete examples of solutions.

Acknowledgments

With thanks to John Hancock, Paul Schofield, Michael Gruenberger, Toyoyuki Takada, Kuniya Abe, Ann-Marie Mallon, Chris Mungall, Shigeharu Wakana, Toshihiko Shiroishi, Yuichi Obata and InterPhenome members for meaningful discussion. This work is supported by the Management Expenses Grant for RIKEN BioResource Center, MEXT.

References

1. Aranguren M.E., Antezana E., Kuiper M., Stevens R.: Applying ontology design patterns in bio-ontologies, Proc. of 16th International Conference LNAI 5268, 7-16. (2008)
2. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening Ontologies with DOLCE, Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, 13th International Conference. 166-181. (2002)
3. Rector A.L., Gangemi A., Galeazzi E., Glowinski A.J., Rossi-Mori A.: The GALEN CORE Model Schemata for Anatomy: Towards a Re-usable Application-Independent Model of Medical Concepts. Twelfth International Congress of the European Federation for Medical Informatics, MIE-94, 229-233
4. Minsky M.: A Framework for Representing Knowledge, in: Patrick Henry Winston (ed.), The Psychology of Computer Vision. McGraw-Hill, New York, (1975).
5. Gkoutos G.V., Green E.C., Mallon A-M, Hancock J.M. and Davidson D. Using ontologies to describe mouse phenotypes, *Genome Biol* 6, R8, (2005)
6. Grenon,P., Smith,B.: SNAP and SPAN: towards dynamic spatial ontology. *Spat. Cogn. Comput.* 4, 69--103. (2004)
7. Masolo C., Borgo C.: Foundational Aspects of Ontologies (FOnt 2005) Workshop at KI (2005)
8. Masolo C.: Founding properties on measurement. Proceedings of the Sixth International Conference (FOIS 2010)
9. Probst F.: Observations, measurements and semantic reference spaces. *Applied Ontology* 3, 63-89, (2008)
10. Mizoguchi R.: Yet Another Top-level Ontology: YATO, Proceedings of the 2nd Interdisciplinary Ontology Meeting, 2, 91-101. (2009)
11. Stevens, S.S. On the Theory of Scales of Measurement. *Science* 103, 677--680. (1946)
12. Mizoguchi, R.: Tutorial on ontological engineering – Part 3: Advanced course of ontological engineering, *New Generation Computing*, OhmSha & Springer, 22, No.2, pp.198-220. (2004)
13. IAO, <http://code.google.com/p/information-artifact-ontology/>
14. Kozaki, K., Sunagawa, E., Kitamura, Y., Mizoguchi, R.: Role Representation Model Using OWL and SWRL, Proc. of 2nd Workshop on Roles and Relationships in Object Oriented Programming, Multiagent Systems, and Ontologies, 39-46 (2007)
15. PATO2YAMATO, http://www.brc.riken.go.jp/lab/bpmp/ontology/ontology_pato2yato.html
16. Smith C.L., Goldsmith C.A., Eppig J.T.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* 6, R7. (2005)
17. Mungall C, Gkoutos G.V., Smith C., Haendel C., Lewis C., Ashburner M.: Integrating phenotype ontologies across multiple species. *Genome Biology*, 11. R2. (2010)
18. IMPC, <http://www.mousephenotype.org/>
19. Masuya H., Makita Y., Kobayashi N., Nishikata K., Yoshida Y., Mochizuki Y., Doi K., Takatsuki T., Waki K., Tanaka N., Ishii M., Matsushima A., Takahashi S., Hijikata A., Kozaki K., Furuichi T., Kawaji H., Wakana S., Nakamura Y., Yoshiki A., Murata T., Fukami-Kobayashi K., Mohan S., Ohara O., Hayashizaki Y., Mizoguchi R., Obata Y., Toyoda T.: The RIKEN integrated database of mammals. *Nucleic Acids Res.* 39, D861-870. (2011)
20. Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L.J., Eilbeck K., Ireland A., Mungall C.J.; OBI Consortium, Leontis N., Rocca-Serra P., Ruttenberg A., Sansone S.A., Scheuermann R.H., Shah N., Whetzel P.L., Lewis S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 25, 1251-1255 (2007)

Annotating Experimental Records Using Ontologies

Alexander Garcia¹, Olga Giraldo², Jael Garcia³

¹University of Arkansas, Biomedical Informatics, Medical Center, Little Rock, Arkansas, USA

²National University of Colombia in Palmira, Valle, Colombia

³Universität der Bundeswehr, Munich, Germany

Abstract. By combining a complex mixture of electronic and paper-based records, researchers carefully document their daily research activities. Although managing such a mixture is a common practice, much information recorded in laboratory notebooks in this manner is often lost for practical purposes. Moreover, lab notebooks are usually disconnected from other information resources that researchers frequently use. Interestingly, although Electronic Notebooks are available, these have not been widely adopted. Here we present our approach to the problem of managing knowledge in Electronic Laboratory Notebooks. We combine elements from the Semantic Web, e.g. ontologies supporting organization and classification, with elements from Social Tagging Availability: www.biotea.ws.

1 Introduction

Here we present a knowledge-based approach to managing laboratory information; it combines elements from the Semantic Web (SW), e.g. ontologies supporting organization and classification, with elements from Social Tagging Systems, e.g. collaboration. We have developed several ontologies supporting the annotation of experimental data for some of the processes routinely run at the Center for International Tropical Agriculture (CIAT) biotechnology laboratory. To identify those processes and practices we analyzed 15 laboratory notebooks together with their corresponding electronic records, e.g. XLS files. We identified data types, metadata, organization, retrieval and sharing strategies, sources of data and information, as well as ontology support for the annotation of laboratory information. Central to our approach is the symbiosis between ontologies and social tagging systems [1, 2]. As ontologies do not fully cover the whole domain, the annotation of laboratory notebooks is achieved by combining ontology-based and user-generated tags; generating in this manner a social network built upon tagged concepts. Our scenarios range from basic laboratory techniques such as PCR, DNA extraction, Electrophoresis, and others, to those involving complex biological phenotype-genotype relationships. We are extending and reusing existing ontologies such

as the Ontology for Biomedical Investigations (OBI) [3], the Chemical Entities of Biological Interest ontology (ChEBI) [4], Plant Ontology (PO) [5], Gene Ontology (GO) [6], Annotation Ontology [7], amongst others.

2 Experimental Records as Knowledge Repositories

For the analysis of laboratory practices and notebooks we closely followed the methodologies proposed by Tabard *et al.* [8] and Garcia *et al.* [9]. We interviewed 10 biologists, analyzed their laboratory notebooks, 15 in total, and electronic records; our field observation went on for a period of six months. Several interviews were held between the elicitor and the researchers; use cases representing processes, practice and involved data were constantly being built; these use cases were the basis for our ontology development as well as for our iterative prototyping.

Researchers pay little attention to the sequence of the information; for instance, notes for long experiments are spread throughout the entire laboratory notebook and stored in several electronic files. Researchers tag in order to establish an organization strategy; however, the tagging strategy is very personal. They also add marginal notes; these were usually comprised of few descriptive words located in visible areas of the corresponding page. It was also observed that the vocabulary

used to tag was significantly overlapping amongst researchers who were working on conceptually closer topics; this trend has previously been reported by Marlow *et al.* [10]. For instance, researchers studying genes involved in drought tolerance share information with those who participate in field studies involving those samples that were genetically modified to be more tolerant to the lack of humidity and water. Researchers also deal with electronic records; they store photos, XLS files, outputs of specialized analysis software, etc. Researchers tend to store and manage their files in their PCs; electronic records also come from LIMSs. The rhetorical structure, and the ontologies related to the components of such structure, is presented in Figure 1.

3 Our Approach: Towards Self-Descriptive Documents

Interestingly, the lack of strategies for organizing and managing knowledge in documents (paper-based and electronic files) had previously been reported, albeit in a different domain, by Paganelli *et al.* [11]. Semantic annotation of features facilitates the self-descriptiveness of documents; the availability of such semantics is key when managing organizational knowledge [12]. Documents should be able to “know about” their own content for automated processes to “know what to do” with them. By delivering ontology-based annotation facilities combined with tagging functionalities researchers are adding that descriptive layer in order to i) speed up information retrieval, ii) facilitate collaboration iii) generate an organization strategy – sometimes mainly understood by the laboratory notebook owner.

We have structured the descriptive layers by reusing and extending existing ontologies. For supporting the annotation within our scenario we have identified three main layers, namely: i) that related to the document itself, ii) the annotation layer, and iii) that related to the experiment. For the document we investigated several metadata standards such as Dublin Core (DC) [13], AgMes [14], AGROVOC and the National Cancer Institute thesauri (NCIt); annotations were structured by means of the AO [7]; experimental information was structured by reusing and

extending biomedical ontologies such as OBI, ChEBI, AGROVOC, PO, and GO. An illustration of our layered approach is presented in Figure 1 and Figure 2.

The AO is structuring the semantic annotation as well as the tags generated by users. In this way we are supporting complex SPARQL queries involving several ontologies, for instance: “*retrieve from the eLabBook the pages tagged by Tim Andrews or Lisa Watson with the tags rice and iron for which there is a LIMS data entry*”. This query involves highly interrelated information, covering aspects related to the pages within the ELN (document), annotation and experimental information.

```
SELECT ?eLabBook ?page
WHERE {
  ?annotation ann:annotates ?page .
  ?annotation pav:createdBy ?user . ?user
foaf:name ?userName .
  FILTER(?userName = "lisa watson" ||
?userName = "tim.andrews").
  ?annotation ao:hasTag ?tag .
  ?tag tags:name ?tagName .
  FILTER(?tagName = "rice" || ?tagName =
"iron")
  ?eLabBook hasPage ?page .
  ?page hasLIMSDataEntry ?lims
}
```

Our document layer provides the ontology for describing the laboratory notebook as well as classes and properties for representing relations with resources such as the LIMS. It has concepts such as “*creator*” (DC), “*investigator*” (NCIt) and “*laboratory procedures*” (NCIt). As researchers store information as they produce it, time is an issue. For instance, researchers may start to record the growth of a plant at intervals of 15 days; this usually also involves taking pictures of the plant, sampling the soil and keeping daily records for atmospheric conditions. The records for his/her observations will be spread all over the laboratory notebook, making it difficult for the researcher to have a unified view of the work. For such situations time stamps are not sufficient; the property “*has_labprocedure*” is practical because this property can be further specialized by “*laboratory procedures*” (NCIt) and “*study subject role*” (OBI_0000097); facilitating thus the interlinking of records so that researchers may retrieve unified views for specific “*laboratory procedures*” combined with “*study subject roles*” – e.g. “*plant structure*” (PO:0009011).

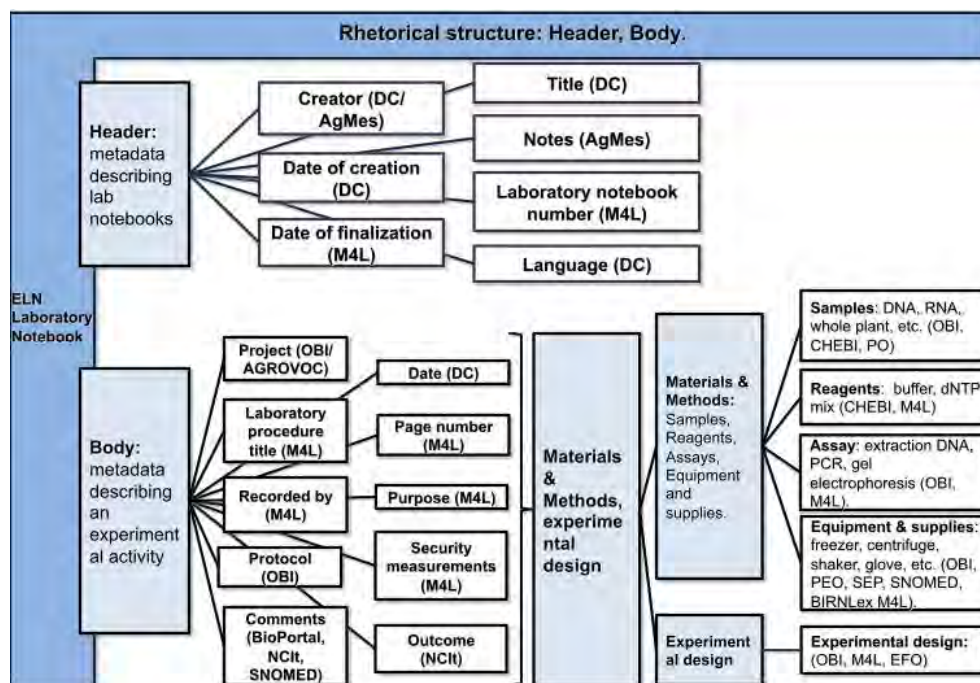


Figure 1. The rhetorical structure

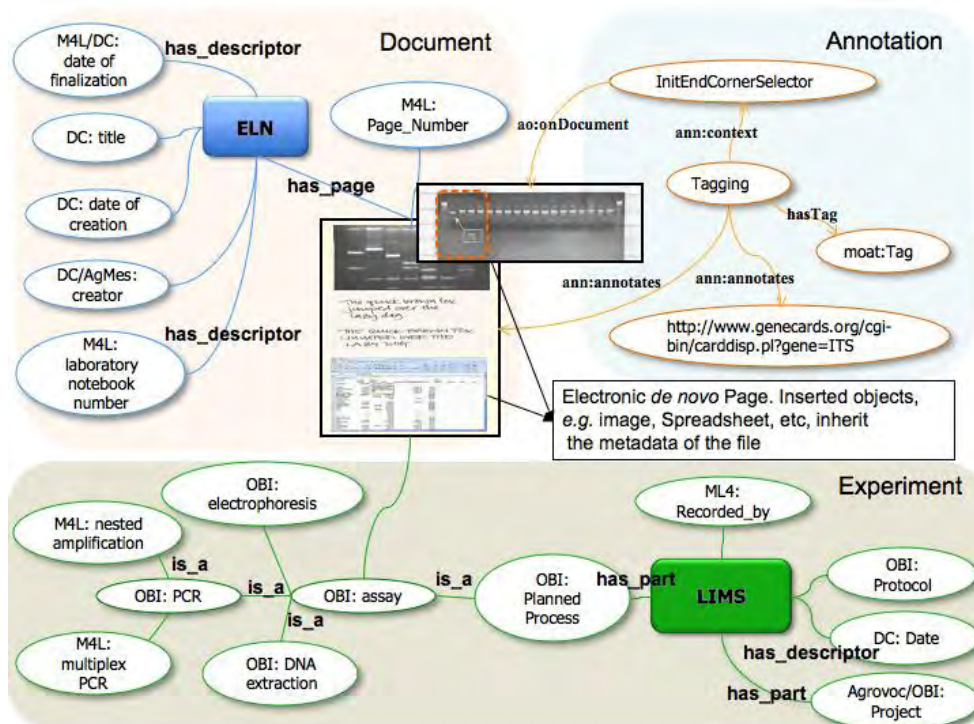


Figure 2. Our layered approach¹

¹ M4L stands for Metadata for Laboratory; it denotes those ontologies developed specifically for our scenarios. M4L ontologies are available at www.biotea.ws

3.1 Structuring the Annotations

For structuring the tagging we decided to use the AO. It was conceived to support the annotation of documents; it can be used to save the tagging data and publish it as Linked Data. Tags are part of the organizational strategy used by researchers in their lab-notebooks; tags are also used to relate specific areas, within pages of the lab-notebooks, to internal and/or external resources. As illustrated in Figure 3, not only is it possible to annotate the entire ELN page, but also a selected portion of it; the AO facilitates the definition of relations between electronic pages and internal/external resources. Also, as part of the example presented in Figure 3, there are two annotations made by the same user; being the user represented by her/his FOAF. Both ANNOT_1 and ANNOT_2 are Qualifier annotations, *i.e.* an ontological term – *GeneBank:AB005238*, is attached to the tag – *Partial sequence on psy promoter*. ANNOT_1 annotates a portion of the image in the ELN and relates it to an ontological term, which is also related to a scientific paper by means of ANNOT_2. In this way it is possible to enrich

the information in the ELN with ontological terms, free text, papers, images, videos, and anything for which there is an URI. Having other type of annotations such as Note, Definition, and Erratum is also possible.

By tagging laboratory notebooks researchers are generating clouds of tags. As laboratory notebooks don't have tables of contents, users identified the clouds of tags as a valuable resource for rapid inspection of contents. By facilitating the generation of tags, combining those coming from ontologies with those provided by users, we are supporting queries such as “*retrieve from the eLabBook those pages having an EXCEL spreadsheet and that have been tagged by Tim Andrews with the tag rice and optionally with PCR*”.

```
SELECT ?eLabBook ?page ?file
WHERE {
  ?annotation ann:annotates ?page .
  ?annotation pav:createdBy <http://www
.tags4lab.org/foaf.rdf#tim.andrews> .
  ?annotation ao:hasTag ?tag .
  ?tag tags:name "rice" .
  OPTIONAL
  {?tag moat:tagMeaning OBI:PCR} .
  ?eLabBook hasPage ?page .
  ?page hasExcel ?file
}
```

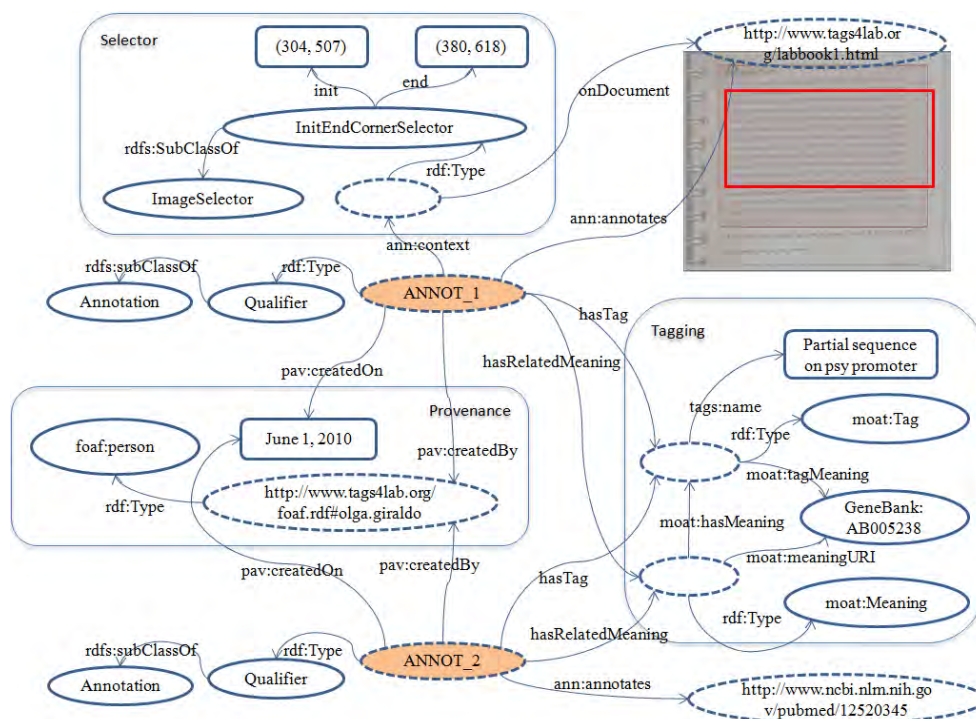


Figure 3. Supporting the annotation of laboratory records with AO

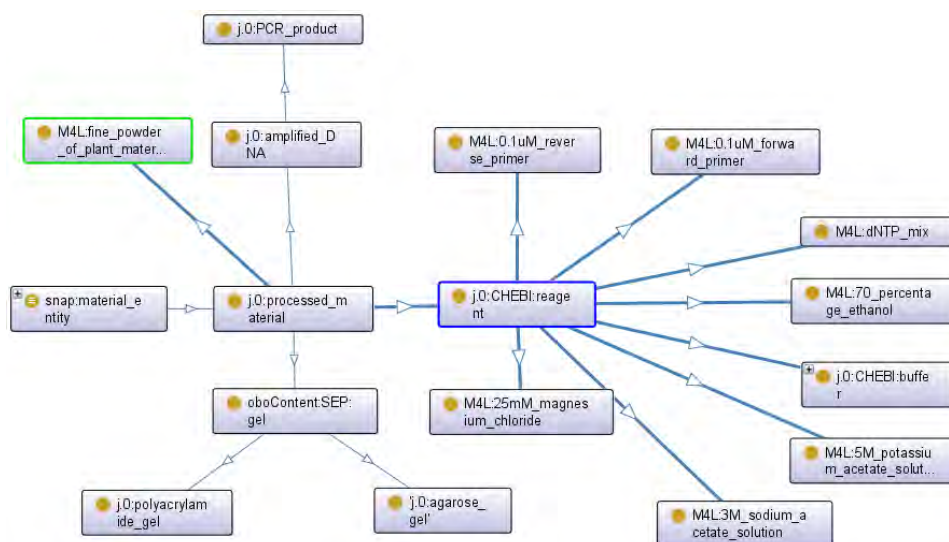


Figure 4. A snapshot of M4L. j.0=undeclared name space.

3.2 Experimental Data

Describing biomedical investigations involves bringing together a wide variety of highly interrelated data. There are various levels of complexity and granularity, combined with a wide range of materials and equipment [15]. Within the structure provided by the Basic Formal Ontology [16] OBI defines an extendible set of terms aiming to describe biological and clinical investigations. For the experimental layer we have followed the structure provided by BFO and OBI (version 2009-11-06, aka version 1.0). OBI defines an “investigation” (OBI_0000066) as a “process” (BFO) that involves, amongst others, the general planning of the “study design” (OBI_0500000), its corresponding execution, the documentation of results and the “interpreting data” (OBI_0000338) so that conclusions can be derived and supported.

Similarly to the clinical domain, processes in plant biotechnology also involve several sub-processes. From BFO, OBI uses “*material entity*” (BFO) as the basis for physical artifacts. Material entity is “*an independent continuant that is spatially extended whose identity is independent of that of other entities and can be maintained through time*”. A material entity is, for instance, a collection of random bacteria, a chair, or the dorsal surface of the body. Material entities can have “*roles*” (BFO); for instance the “*study subject role*” (OBI_0000097). “*Functions*” depend on the

design or physical structure of the entity; for instance, “*measure function*” (OBI_0000453), “*freeze function*” (OBI_0000375). Functions are considered inhered, “*inheres in*” (BFO) by the material entity and “*realized by*” (BFO) the “*role*” that is assumed by the “*material entity*” within the process.

As illustrated in Figure 4, we are reusing and extending OBI in combination with other ontologies so that our use cases are fully covered; these have been selected from those laboratory procedures, “*assay*” (OBI_0000070), commonly carried out by the Biotechnology group at CIAT; a snapshot of M4L is presented in Figure 4. To illustrate some of the experimental ontologies that have been developed we selected the small scale extraction of high quality DNA “*assay*” (OBI_0000070). Three planned processes are part of this assay, namely: harvesting the plant material, pulverizing it, and extracting the DNA.

3.2 Sample Preparation for Assay and DNA Extraction

Both, harvesting the plant material and pulverizing it, illustrate the “*sample preparation for assay*” (OBI_0000073) class from OBI. Initially researchers use “*scissors*” (SNOMED-CT ID 64973003), for which there is a “*mechanical function*” (OBI_0000379) to obtain the “*juvenile leaf*” (PO:0006339) or an “*adult leaf*” (PO:0006340). The “*leaf*”

(PO:0009025) assumes the “*study subject role*” (OBI_0000097). The “*leaf*” is stored in a “*reclosable bag*” (M4L) that is stored in a “*portable ice chest*” (M4L). The vegetal material is then stored in a “*freezer*” (PEO, “*freeze function*” OBI_0000375). Pulverizing the “*leaf*” starts by taking the “*leaf*” out of the “*freezer*”; the “*frozen leaf*” (M4L) is a “*material entity*” (snap:MaterialEntity). This material is then converted into a “*fine powder*” (M4L); such “*fine powder*” is a “*processed material*” (OBI_0000047).

A “*microcentrifuge tube*” (M4L) is then used to store the “*fine powder*” (M4L). “*DNA*

extraction buffer” (M4L), the buffer’s “*role*” is to dissolve the tissue; facilitating in this manner the extraction of the DNA. The whole process, from the “*fine powder*” is illustrated in Figure 5. The DNA extraction ontology has over 140 classes; we are reusing BFO properties such as “*inheres in*” (BFO_0000052), “*bearer of*” (BFO_0000053), “*realized by*” (BFO_0000054), and “*realizes*” (BFO_0000055). Other sources for relationships come from the Relation Ontology [17]. Twenty-nine fully documented new terms, from M4L, have been added to the OBI structure that we are reusing; terms from other ontologies are also being reused.

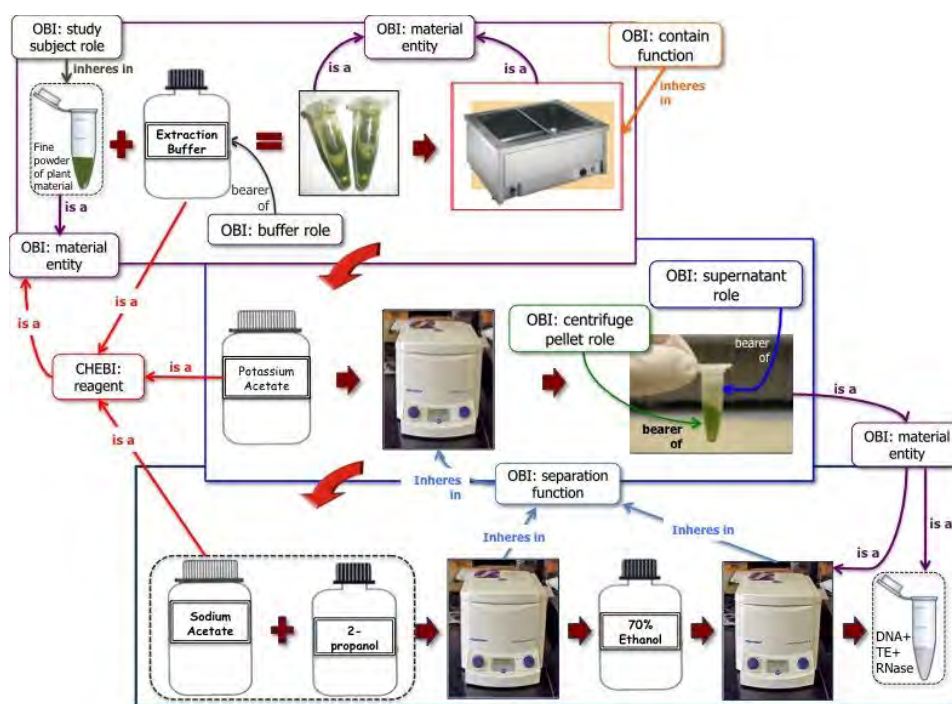


Figure 5. Extracting DNA

4 Discussion

Biologists have so far managed to balance and combine paper-based records with an intricate network of electronic records (LIMS, spreadsheets, URLs, etc). However, as it is necessary to process more and more data in a systematic interoperable manner, the information management infrastructure will have to facilitate data capture/processing in an integrative way; more importantly, the semantic layer within these infrastructures needs to be vastly improved. Such layer requires interoperable, well modularized

ontologies rather than monolithic ones. Although several ELNs have been proposed [18, 19], and replacing paper-based records has been a consistent trend for several years, the technology has not yet been widely adopted [18]; Laboratory Information Management Systems (LIMS) in combination with paper-based laboratory notebooks continue to be commonly used; particularly in academic environments [8, 20].

Sharing and organizing information happens on a concept basis; for instance, researchers studying genes involved in iron transport share information with those who

undertake nutritional studies assessing the effects of iron intake in human populations. Such concept-based folksonomy was easily observed; ontologies supporting the annotation of laboratory records and practices made it easier for researchers to share and interact. By the same token, being able to tag with user generated tags as well as ontology-based tags, facilitated the organization of information. Interestingly, although tagging practices were personal, these were similar amongst those researchers working on conceptually similar projects. Tags were also a valuable resource providing new terms for our ontologies.

5 Conclusions and Future Work

Using ontologies to support the annotation of experimental activities requires highly interoperable ontologies. Although there is a generic extensible ontology for relations in the biomedical domain, not all of them are actively using it. Also, although biomedical ontologies are pursuing a thoughtful and important ontology development standardization effort; there are still methodological gaps. OBI facilitates the description of biomedical experiments; such effort implies interoperating with other ontologies; from our experience interoperability between OBI and OBO ontologies was not a straightforward process. Standardization efforts based on minimal amounts of information, grounded in existing ontologies, are important for facilitating interoperability across laboratories; such efforts should focus on providing easily implementable data capture templates.

We have presented a semantic layered approach that facilitates the self-description of experimental records from the Biotechnology laboratory at CIAT. We have reused CHEBI, OBI, BIRNLex as well as other ontologies; within our OBI scaffold we have added 145 terms, all of them extracted from our experimental records. We envision a paperless laboratory in which Ubiquitous Computing takes advantage of SW technology, for supporting knowledge management, and folksonomy principles for facilitating the collaboration. We have started by making extensive use of ontologies for supporting knowledge management, by the same token we are facilitating interaction in similar ways to

those currently available in social networks; our interaction is based upon research activities and concepts. In the near future we will improve the usability of our prototype, we are also planning to release the software to the open source community; we are currently continuing with our ontology development effort.

References

1. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: *Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering*. International World Wide Web Conference - Workshop on Social and Collaborative Construction of Structured Knowledge (CKC), Canada (2007)
2. Almeida, A., Sotomayor, B., Abaitua, J., López-de-Ipiña, D.: *Folk2Onto: Bridging the gap between social tags and ontologies*. European Semantic Web Conference, Tenerife, Spain (2008)
3. Brinkman, R., Courtot, M., Derom, D., Fostel, J., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Soldatova, L., Stoeckert, C., Turner, J., Zheng, J., consortium, t.O.: *Modeling biomedical experimental processes with OBI*. *Journal of biomedical semantics* **1** (2010) S7
4. de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C.: *Chemical entities of biological interest: an update*. *Nucleic Acids Research* (2009)
5. Pankaj, J., Shulamit, A., Katica, I., Elizabeth, A., Kellogg, S., Susan, M., Anuradha, P., Leonore, R., Seung, Y., Rhee, M., Martin, S., Mary, S., Lincoln, S., Peter, S., Leszek, V., Doreen, W., Felipe, Z.: *Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages*. *Comparative and Functional Genomics* **6** (2005) 388-397
6. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel, T.L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G.: *Gene Ontology: tool for the unification of biology*. *The Gene Ontology Consortium*. *Nature Genetics* **25** (2000) 25-29
7. Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., Clark, T.: *An open annotation ontology for science on web 3.0*. *Journal of biomedical semantics* **2** (2011) S4
8. Tabard, A., Eastmond, E., Mackay, E.W.: *From Individual to Collaborative: The Evolution of*

- Prism, a Hybrid Laboratory Notebook. Computer Supported Cooperative Work. ACM, San Diego, California, USA (2008)
9. Garcia, A., O'Neill, K., Garcia, L.J., Lord, P., Stevens, R., Corcho, O., Gibson, F.: Developing Ontologies within Decentralised Settings. In: Chen, H., Wang, Y., Cheung, K.-H. (eds.): Semantic e-Science, Vol. 11. Springer US (2010) 99-139
 10. Marlow, C., Naaman, M., Boyd, D., Davis, D.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. 7th Conf. Hypertext and Hypermedia. ACM Press, Edinburgh, Scotland, United Kingdom (2006) 31-40
 11. Paganelli, F., Pettenati, M.C., Giuli, D.: A Metadata-Based Approach for Unstructured Document Management in Organizations. Information Resources Management Journal **19** (2006) 22
 12. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Journal of Web Semantics **4**(1) (2005) 15
 13. <http://dublincore.org/>: Dublin Core Metadata Initiative (2010)
 14. <http://aims.fao.org/en/agmes-metadataset/>: AgMes metadata element set. (2010)
 15. Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Soldatova, L.N., Stoeckert Jr., C.J., Turner, J., Zheng, J., consortium., t.O.: Modeling biomedical experimental processes with OBI. Journal of Biomedical Semantics **1** (2010) 11
 16. Grenon, P., Smith, B., Goldberg, L.: Biodynamic Ontology: Applying BFO in the Biomedical Domain. Ontologies in Medicine (2004) 20-32
 17. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in Biomedical Ontologies. Genome Biology **6** (2005) R46
 18. Van Eikeren, P.: Intelligent Electronic Laboratory Notebooks for Accelerated Organic Process R&D. Organic Process Research & Development **8** (2004)
 19. Talbott, T., Peterson, M., Schwidder, J., Myers, J.D.: Adapting the electronic laboratory notebook for the semantic era. In: McQuay, W., Smari, W.W., Kim, S.-Y. (eds.): International Symposium on Collaborative Technologies and Systems. IEEE Computer Society (2005)
 20. Butler, D.: Electronic notebooks: A new leaf. Nature **436** (2005) 20-21

Ontology Driven Data Collection for EuPathDB

Jie Zheng, Omar S. Harb, Christian J. Stoeckert Jr.

Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA, USA

Abstract. EuPathDB is a public resource of protozoan parasite genomic and functional genomic data. To address community needs, information on isolate specimens, and on genetic manipulation and phenotype, data will be collected directly from scientists. In order to facilitate data exploration, exchange, sharing and reuse, such data needs to be well-structured with standardized annotation. However, data collection in a uniform format remains challenging. In this report, we leverage existing ontologies to semantically represent the two cases of (1) isolate and (2) genetic manipulation and phenotype data with a focus on the needs/requirements of the EuPathDB community. Using ontology-based models, we designed submission forms and incorporated ontology terms for annotation with the goal of minimizing the burden on end users to submit standardized data.

Keywords: Ontology for Biomedical Investigation (OBI), ontology, EuPathDB

1 Introduction

Protozoan parasites are a major cause of global human and veterinary infectious diseases, such as malaria, toxoplasmosis, cryptosporidiosis, Chagas disease, sleeping sickness and leishmaniasis. Unfortunately current treatments are limited due to the rise of parasite drug resistance and immune evasion. The Eukaryotic Pathogen Database (EuPathDB; <http://eupathdb.org>) project integrates genomic and functional genomics data from over 30 different protozoan parasite species [1]. EuPathDB also integrates parasite isolate data, and genetic manipulation with resulting phenotype data but in a limited manner due to the heterogeneity of what is currently obtained. Desired information includes the geographic location of parasite isolate specimens collected, pathogen host information, and genetic manipulation and phenotype information associated with specific genes as these are important for parasite epidemiology, and vaccine anti-parasitic drug research. Currently, EuPathDB integrates isolate data from GenBank [2] which in turn, accepts sequence data and associated information directly from individual researchers. The submission to GenBank has clear requirements for sequence format and annotation. However, the completeness of information associated with sequence data is

mainly dependent on the submitters. For example, when submitting isolate data to GenBank, there are no requirements for providing the parasite's host information, such as host organism, clinical information, and geographic location from where the isolate specimens were collected. This constitutes a big hurdle for EuPathDB and requires manual intervention to standardize, for example, the terms used for organism and geographic location. In addition, some important isolate information is lost during integration into EuPathDB due to its deposition as free text in the GenBank records. EuPathDB has some semi-structured genetic manipulation and phenotype data for *T. brucei*. Technically, genetic manipulation and phenotype data can also be collected through submission of User Comments available on EuPathDB gene pages. However, while such free text contributions are invaluable they are less useful for data exploration. For data integration, sharing and reuse, there is a need to collect isolate, genetic manipulation and phenotype data in a well-structured manner with standardized annotation.

Usage of the Gene Ontology (GO) [3] by model organism databases and many other resources has led to great success in data exchange, sharing, reuse and analysis. EuPathDB has also utilized GO and related ontologies (e.g., Sequence Ontology (SO) [4]) to

integrate genomic data annotation. The application of these ontologies enables EuPathDB to issue complex queries, such as retrieving functional genomic data across multiple species based on orthology [1].

The Ontology for Biomedical Investigations (OBI) is being developed for supporting consistent annotation of biological and clinical investigations [5]. It covers the terms to describe all aspects of an investigation including biological materials, protocols, generated data and type of analysis applied to the data. OBI is based on the Basic Formal Ontology (BFO) and follows Open Biomedical Ontologies (OBO) Foundry principles [6]. OBI is interoperable with other biomedical ontologies under the OBO Foundry umbrella since they are built on the basis of a common top-level ontology, BFO, and use a common set of relations. Each OBO Foundry ontology covers terms in a specific domain. For example, OBI focuses on experimental processes, GO is used for gene and gene product annotation and contains three main components, cellular component, molecular function, and biological process [3], and the Phenotype And Trait Ontology (PATO) defines terms for the description of phenotype and quality [7].

In this report, we describe how we apply OBI to model isolate and genetic manipulation with resulting phenotype data, generate effective submission forms, and provide terms for annotation from recommended OBO Foundry ontologies including OBI. Our experience demonstrates that starting with a semantic framework such as OBI is an effective approach of creating a well-structured data collection form.

2 Method

The following steps were applied to generate the submission form for collecting data from individual investigators.

2.1 Semantically Represent Data For Collection Using OBI

This step involved identifying the categories of data and information required to sufficiently characterize isolate specimens or genetically modified parasite phenotypes, followed by OBI-based modeling of the defined categories to

capture their interconnectedness and relationships.

2.2 Generate the Submission Form Based on the Ontology Model

Using the ontology model as a guide, categories from step 1 (section 2.1) were organized logically with related categories grouped together in the two forms. For the isolate submission form, category order was as follows: isolate information (species, type, etc.), location of collected sample (geographic location, country, province, city), isolation source (host organism sample, or environmental sample), and nucleotide sequence information (sequence name, type, sequence). The genetic manipulation and phenotype submission form is organized as two main sections: genetic manipulation including genetic modification method, markers and/or reporters used in the modification; and assays used for investigation of the impact on the organism or the cellular location, molecular function, or biological process associated with the gene product.

We then determined which OBO library ontologies to use for various types of data and identified the terms in an ontology needed for standardized annotations. Some types of data, such as organism species and country, have a long list of ontology terms. In this case, we provide commonly used terms by surveying existing datasets.

The choice of format for the submission forms was based on feedback from EuPathDB end users with experience with these types of datasets.

3 Result

We first describe the data to be collected and then describe the generation of the submission forms for acquiring isolate data and genetic manipulation and phenotype information.

3.1 Isolate Submission Form

Parasites are isolated from samples and typed by their contained nucleotide sequences. This data and associated meta-data are important for epidemiological research related to parasite spread and resistance. The meta-data include information specific for the isolate specimen (species in the isolate, geographic location,

collection source), and host organism specific information (species, age, and clinical information). We describe data of interest and their relations using OBI and other ontologies.

The graphical representation in Fig. 1 allows effective and rapid identification of connections and relationships between the various items in the submission forms. For example, viewing the middle left section of the graphic illustrates how an isolate specimen is a subclass of *specimen*, while specimen ID is a *symbol* used to uniquely identify the isolate specimen studied. The specimen encompasses the target species of interest (isolate) which is a subclass of *organism* (bottom left of Fig. 1) and other microorganisms and is the output of *specimen creation process* (middle left of Fig. 1). Moving up the graphic in Fig. 1, the input of the *specimen creation process* is either the environmental source or host material (both subclass of *material entity*) that is *located in* a specific *geographic location*. In turn, the host material (upper middle of Fig. 1) is *part of* a host *organism* that *has qualities* such as *sex*, *age*, and *symptoms* (upper right of Fig. 1). The sequencing type assay (center of Fig. 1) is a subclass of *assay* and is linked to the isolate specimen as a *specified input* and its *specified output* is nucleotide *sequence data* for DNA from the isolate. The *textual entity*, locus/product name and sequence description (lower middle of Fig. 1), specify the *sequence*. The GenBank ID (bottom right of Fig. 1) is a subclass of *symbol* that is used to identify the *sequence* submitted to GenBank.

Submission of sequence data to GenBank and obtaining a GenBank ID(s) are essential steps for scientists sharing their data directly

or via publications. Hence, it is critical that our submission form provides a convenient method to make data available in both GenBank and EuPathDB. In fact, enabling easy submission to GenBank through use of the form is a major incentive for community use. Therefore, essential to developing our submission form is developing a parser that generates GenBank “ready” files yet retains the added structure and standardized terms. It is anticipated that although EuPathDB will initially collect isolate data with the form, the data will first be delivered to GenBank to ensure it is properly archived and tagged with GenBank IDs. Subsequently, isolate data will be retrieved from GenBank for integration into EuPathDB and querying through its web sites.

A spreadsheet (Excel) format was used for the isolate specimen submission form because such data are generally already collected in this format according to our community advisors familiar with this data type. The spreadsheet format also facilitates submission of multiple nucleotide sequences associated with the isolates which can be cumbersome on typical web-based forms.

Ontology terms that can be used in the data annotation are provided as drop-down lists in the submission form. These include, OBO ontologies, NCBI Taxon, OBI, Environment Ontology (EnVO), PATO, Ontology for General Medical Science (OGMS) and Gazetteer (GAZ). Having standardized annotation such as these in GenBank will greatly reduce or eliminate the need for manual intervention to integrate isolate data from GenBank into EuPathDB.

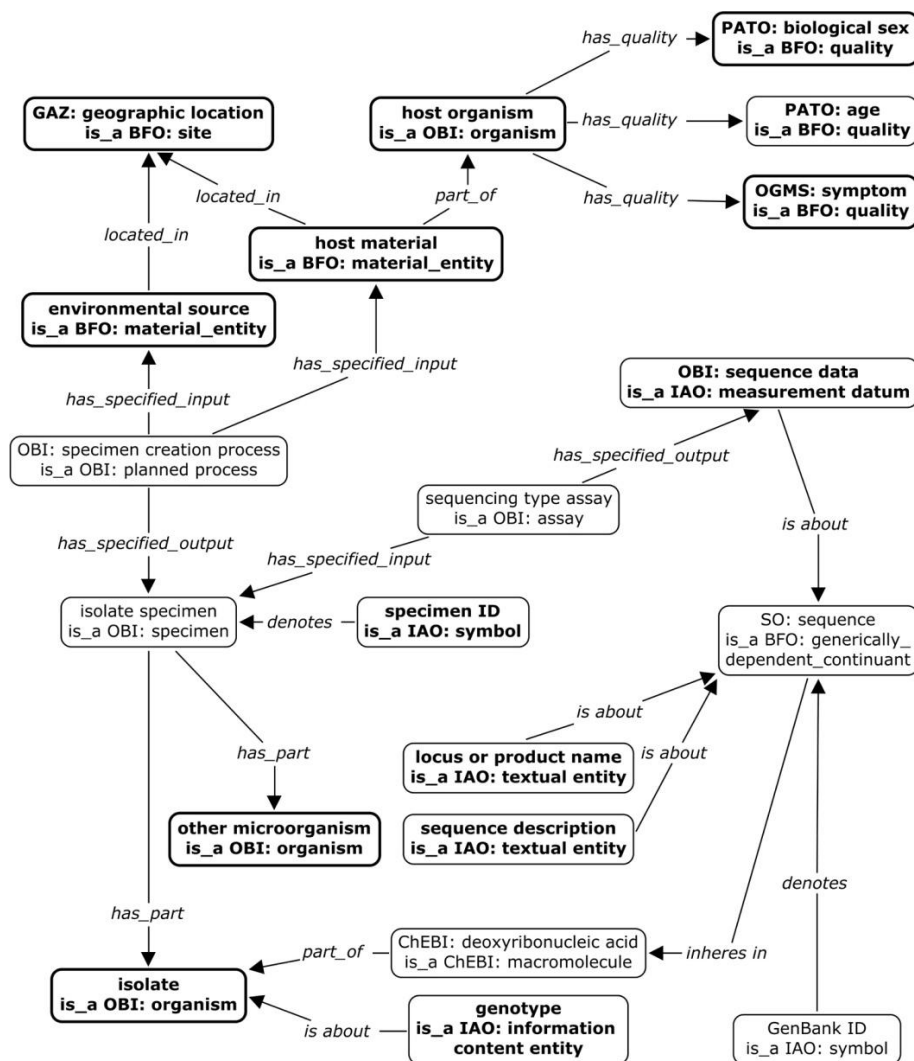


Figure 1. Ontology-based representation of sequence related isolate data. The ontology terms are indicated by using ontology name abbreviation as prefix. Italicized text represents relations. The data collected in the submission form are in the bold font. The fields require ontology terms are in thick border box. IAO stands for Information Artifact Ontology. ChEBI stands for Chemical Entities of Biological Interest ontology.

3.2 Genetic Manipulation and Phenotype Submission Form

Observed phenotype(s) that can be linked to specific genetic modifications are valuable for the development of novel anti-parasitic drugs. In addition, knowledge of when in the parasites life cycle an observed phenotype occurs and has an effect is important. To this end, we represent the data of interest using ontologies (mainly OBI and GO), as illustrated in Fig. 2.

Phenotype data may refer to the impact of gene modification on four possible observed features:

- Quality of the organism

- Cellular location of gene product
- Molecular function of gene product
- Biological process of gene product

The genetically modified parasite is a subclass of *genetically modified organism* and is generated by *genetic transformation* process (top section of Fig. 2). The left and lower sections of Fig. 2 illustrate that *assays* are performed to examine the genetically modified parasite about *organismal quality* (eg. viability), *cellular component* the gene product is *located in*, effects on its *molecular function*, or *biological process* it *participates in* at specific *lifecycle stage*.

The OBI terms will be used in the genetic modification method and assay fields. The GO terms will be listed in cellular component, molecular function and biological process fields and Ontology for Parasite Lifecycle (OPL) will be used for annotation of lifecycle stages.

Driven by our communities' needs genetic manipulation and phenotype submission will occur using a web form since EuPathDB plans to collect these data directly from specific locations at the web site (Gene pages containing information that can be linked to genetic manipulation and phenotype). One big advantage of a web form is that it can change dynamically based on the users input. For example, not all kinds of genetic modification methods use selectable markers and/or reporters. Once a user selects a specific

modification method like gene knock out or gene knock in, the form dynamically displays questions about marker(s) and reporter(s). The ontology terms used for annotation are also changed depending on the input data that can short the term list. For example, different parasites may have different lifecycle stages. A subset of lifecycle stage terms will be listed based on which kind of parasite the gene is derived from. All these can minimize the efforts of the end users during the submission process.

We have collected comments on the submission forms from some of the end-users. The forms have been adjusted based on the feedback. The isolate submission has been approved by the end users and will be distributed to more parasite researchers.

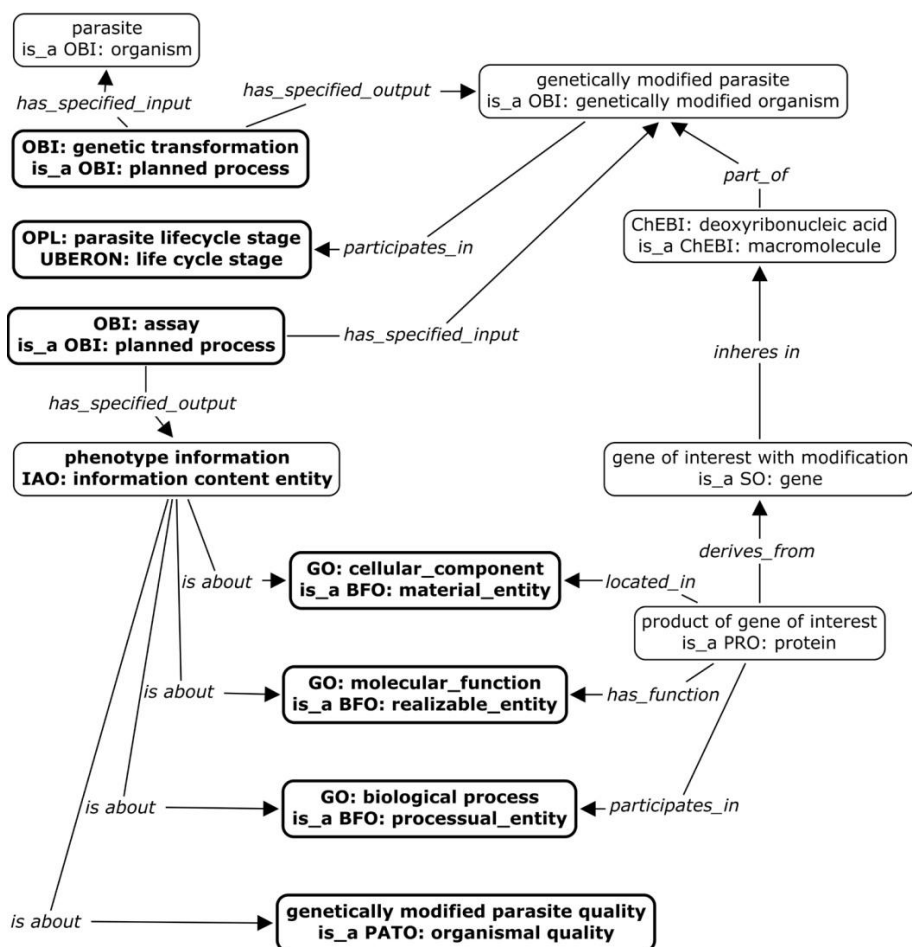


Figure 2. Ontology-based representation of phenotype data. The ontology terms are indicated by using ontology name abbreviation as prefix. Italicized text represents relations. The data collected in the submission form are in the bold font. The fields require ontology terms are in thick border box. IAO stands for Information Artifact Ontology. ChEBI stands for Chemical Entities of Biological Interest ontology.

4 Discussion

Consistently-applied annotations will facilitate data integration, sharing, and (re)analysis. In recent years, ontologies have been widely used in supporting the consistent annotation of the biological data. However, it is still a big challenge to collect standardized data directly from scientists. Complicated submission forms and the use of ontology terms for annotation often inhibit scientists from submitting their data. However, collection of low quality data is less useful for sharing, exchanging and analysis. Our goal is to minimize the efforts of the submitters and capture important data with standardized annotation at same time. In this report, we apply a semantic framework to submission form designs. It is an efficient approach to organize the data in a logical fashion and to identify appropriate bio-ontologies to be used in annotation.

The isolate form will allow scientists to submit multiple sequences to GenBank without going through the GenBank submission process which can be a challenge to scientists especially when submitting many isolate records. Using the genetic manipulation and phenotype collection form to capture crucial genetic modification and phenotype data using ontology terms will facilitate the ability of EuPathDB to share and exchange data with other genetically modified parasite phenotype databases, such as RMgmDB (<http://pberghel.eu/>). Furthermore, high quality isolate and genetic modified parasite related phenotype data, will enable users of EuPathDB to perform integrated searches of the types: "Compare sequence data from *Plasmodium* isolates that are restricted to East Africa to those from West Africa", "List genes that when knocked out result in a defect in parasite growth during the intraerythrocytic cycle", "List genes fused to green fluorescent protein (GFP) that when expressed are located in the cell membrane".

Currently the submission forms are in the prototype stage. We will distribute the isolate submission form to EuPathDB user communities and incorporate the genetic manipulation with associated phenotype form

into EuPathDB websites. Based on user feedback, the forms and underlying ontology-model will be improved to achieve the goal of collecting critical information from individual scientists about their experiments that is structured yet is not burdensome, and provides incentives (such as facilitated submission to the GenBank archive).

Acknowledgments

We thank Dr. G Robinson, Dr. R Chalmers, Dr. CJ Janse, Dr. G. Widmer, Dr. L. Xiao, and Dr. SM Khan for their valuable comments on the submission forms.

This research is supported by NIH grant 5R01GM93132-1 and by the National Institute of Allergy and Infectious Diseases at the National Institutes of Health Award NO1-AI900038C Contract No. HHSN272200900038C.

References

1. Aurecochea C., Brestelli J., Brunk B.P., Fischer S., Gajria B. et al.: EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Res* 38(Database issue): D415-419 (2010)
2. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W.: GenBank. *Nucleic Acids Res* 37(Database issue): D26-31 (2009)
3. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H. et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1): 25-29 (2000)
4. Eilbeck K., Lewis S., Mungall C.J., Yandell M., Stein L., Durbin R., Ashburner M.: The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* 6:R44 (2005)
5. Brinkman R.R., Courtot M., Derom D., Fostel J.M., He Y. et al.: Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 1 Suppl 1: S7 (2010)
6. Smith B., Ashburner M., Rosse C., Bard J., Bug W. et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25(11): 1251-1255 (2007)
7. Mungall C.J., Gkoutos G.V., Smith C.L., Haendel M.A., Lewis S.E. et al. Integrating phenotype ontologies across multiple species. *Genome Biol* 11(1): R2 (2010)

Developing an Application Ontology for Biomedical Resource Annotation and Retrieval: Challenges and Lessons Learned

Carlo Torniai¹, Matthew Brush¹, Nicole Vasilevsky¹, Erik Segerdell¹,
Melanie Wilson¹, Tenille Johnson², Karen Corday², Chris Shaffer¹, Melissa Haendel¹

¹Oregon Health & Science University, Portland, OR, USA

²Harvard Medical School, Boston, MA, USA

Abstract. The eagle-i project has been developing a semantic search portal for biomedical research resources. A unique feature of eagle-i is that the data collection and search tools are completely driven by ontologies. This has been a source of challenges and opportunities regarding use of biomedical ontologies in real-world applications. In this paper, we address our approach and lessons learned for balancing practical project requirements for design and implementation of an ontology driven application, with a desire to conform to best practices for biomedical ontology development.

Keywords: biomedical ontologies, application ontology, resource, eagle-i, ontology reuse.

1 Introduction

An important challenge in biomedical research is the ability to find relevant scientific resources such as reagents, instruments, and protocols, thereby reducing time-consuming and expensive duplication of resource development. For resources that are publicly available, information connecting them to relevant organisms, genotypes, genes, site of action, and other key biological search facets is frequently not available within public databases or from company catalogs. Furthermore, there exist numerous resources that are not published in journals or listed on websites. The eagle-i Consortium (www.eagle-i.org/home) aims to help researchers find biomedical research resources. The goal of eagle-i is to collect data about these “invisible” resources from the labs that use them, and make this information available through a semantic search portal. eagle-i also aims to be interoperable with and contribute to similar ontology-based efforts to represent publicly available research resources in repositories such as the Neuroscience Information Framework (NIF) [1] and the Resource Discovery System (RDS) [2]. These efforts have the main benefit of allowing linkage to very many data sets for gene function, expression, phenotypes, biological pathways, etc.

Use of interoperable ontologies will enable understanding of the experimental context in which these data are collected, and support new hypothesis generation.

The architecture of the eagle-i system includes four main components: institutional triple-store repositories; a federated network; a data collection tool, and a central search application. In order to support semantic retrieval of resource data, the underlying data model is based on a modular set of ontologies. A unique feature of the project is that the user interface and logic of both the data collection and search tools are driven by ontologies, allowing these applications to seamlessly change in response to data-driven ontology enhancements [3]. In this paper, we present our approach and lessons learned in the process of developing the eagle-i ontology modules in compliance with project requirements, best practices for biomedical ontology development, and interoperability with other ontology-based resource systems and community ontologies.

2 Modeling Approach

Our modeling approach had three main drivers. The first was to represent real data collected about resources. The second was to have the ontology control the user-interface (UI) and the

logic of the data collection tool and search application. The third was a commitment to build a set of ontologies that could be reusable and interoperable with other ontologies and existing efforts for representing biomedical entities. This latter requirement translated into decisions to a) follow OBO Foundry [4] principles and best practices for biomedical ontology development and b) engage in active discussions within the bio-ontology community in order to provide context for eagle-i interoperability and align with domain-wide standards for resource representation (<http://bit.ly/rrcoord>).

We began our modeling effort by collecting a preliminary set of data with the goal of identifying key properties for each resource type collected by eagle-i. These include reagents, instruments, services, model and non-model organisms, protocols, biospecimens, human studies, and research opportunities. We then asked the eagle-i team to identify a set of queries relevant to each of these resources. For example, “Which laboratories in the United States are equipped with high-resolution ultrasound machines for brachial artery reactivity testing (BART)?” or “Find *in situ* hybridization protocols for whole-mount preparations of *Aplysia*.” Over 300 queries were generated and analyzed to first identify their relevance and semantic linkage, and then to specify the relations required to answer these queries. Using the preliminary data and the analysis of the queries, we defined a preliminary high-level data model. Next we identified a set of classes and properties from existent biomedical ontologies that could be reused to implement this model, as well as those that had to be created *de novo*. Based on data and functional requirements gathered throughout the project, we expanded on this initial ontology in an iterative approach that involved collaboration with NIF, RDS, Ontology for Biomedical Investigation (OBI) [5], and VIVO

[6]. Adherence to our three drivers presented interesting challenges and trade-offs when it came to implementation, which is discussed in the next section.

3 Implementation

3.1 Ontology Reuse and Development Practices

We implemented the eagle-i ontology modules in the Web Ontology Language (OWL) [7] to comply with the *de facto* standard for ontology representation and to exploit its reasoning capabilities. Upon analysis of existing ontologies, we came to the conclusion that those showing the most promising high-level classes and design principles for resource representation belonged primarily to the OBO Foundry constellation. Because of our choice to reuse portions of certain OBO ontologies (OBI [5], Uberon (<http://bit.ly/ubernat>), the Gene Ontology (GO; <http://www.geneontology.org>), the Software Ontology (SWO; <http://www.ebi.ac.uk/efo/swo>), the NIF Standard Ontology (NIFstd), Biomedical Resource Ontology (BRO) [2], etc.), we chose to follow several key OBO Foundry principles for ontology development and reuse. These included:

- Adoption of the Basic Formal Ontology (BFO) [8] as upper level ontology and the Information Artifact Ontology (IAO) [9] for representing ontology metadata.
- Use of the Relation Ontology (RO) [10] for basic properties.
- Adherence to the Minimum Information to Reference an External Ontology Term (MIREOT) [11] principle to reference terms and axioms already defined in other ontologies. MIREOT is a standard whereby a subset of classes and related axioms can be referenced from an ontology, without importing the whole source ontology.

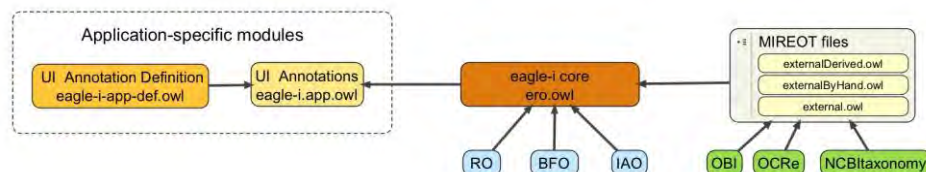


Figure 1. The eagle-i layered, modular ontology structure.

The core module imports BFO, IAO, and RO in their entirety, as well as files containing portions of external ontologies (MIREOT; shown are portions of OBI, Ontology of Clinical Research (OCRe), and the NCBI taxonomy). The application-specific modules define properties, instance values for annotation properties, and property and class annotations used by the eagle-i applications.

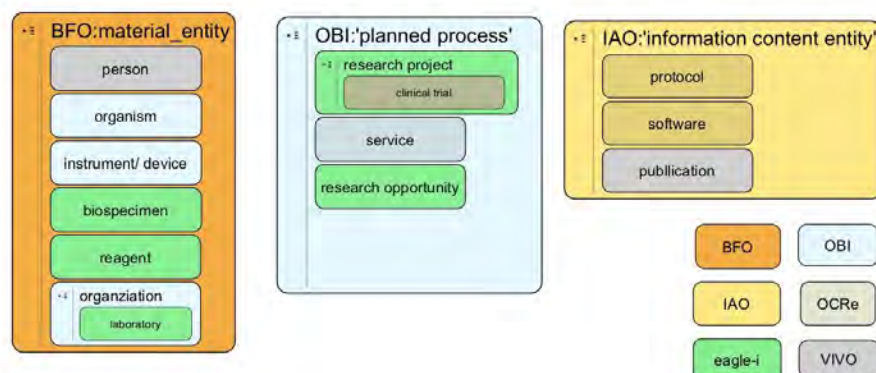


Figure 2. Research resources represented in eagle-i.

The classes from external ontologies referenced through MIREOT are as follows: 507 from OBI, 182 from NCBI Taxon, 53 from SWO, 20 from VIVO, 19 from OCRE, 13 from BRO.

3.2 Layered eagle-i Ontology Modules

We used an approach in which the representation of research resource data is decoupled from the representation of application-specific data used to control the appearance and behavior of the tooling UI. The result is a collection of ontology modules that are encoded and maintained separately, but assemble into a single ontology to support the eagle-i data collection tool and search application (Fig. 1).

The eagle-i core Ontology. The eagle-i core module contains classes and properties used to represent, logically define, and retrieve biomedical research resources (Fig. 2). The eagle-i core has its own unique namespace identifier (ERO) as required by the OBO Foundry. The eagle-i core module imports some external ontologies in their entirety, as well as a set of files containing individual classes and properties collected from external ontologies using MIREOT (Fig. 2). As of April 2011, the eagle-i core module contains 1059 ERO classes (excluding classes imported directly or referenced with MIREOT), 56 object properties, and 59 data properties. The current version of the core ontology under development is available at <http://bit.ly/eagle-i-onto>.

The application-specific modules drive application functionality. The eagle-i application-specific modules contain all the properties and classes required to drive the UIs of the data collection tool and the search application. These are primarily annotation

properties that tell the data and search tools how to display and interact with the ontology classes and properties to which they are attached.

The basic design principle is to define a set of annotation properties and possible instance values for these properties in a 'UI Annotation Definition file' (eagle-i-app-def.owl). For example, the '*inClassGroup*' and '*inPropertyGroup*' annotation properties are used to tag specific classes and properties, respectively, as exhibiting certain application-related features or behavior. Table 1 shows some of the possible instance values for the '*inClassGroup*'¹ and '*inPropertyGroup*' properties, Table 2 describes additional properties defined in the UI Annotation Definition file, and Fig. 4 illustrates how they control various aspects of the data collection tool UI. A second module, the 'UI Annotations file' (eagle-i-app.owl), holds the actual annotations made on core eagle-i classes and properties using these annotation values. These two application-specific modules have a different namespace than the core ontology, and class and property URIs are not numerical since they are not meant to be shared or reused.

¹ Throughout this text, *italics* are used to indicate a term denoting an ontology class, instance or property.

Instance Label	Description	Example
resource	Denotes classes for which instances are collected	<i>'instrument', 'biospecimen', 'protocol'</i>
referenced class	Denotes non resource classes that are used to populate drop down menus in the UI	<i>'technique', 'disease'</i>
data model exclude	Denotes classes or properties that are not included in the model used for the data tool or the search tool UIs	BFO classes such <i>'continuant'</i> or <i>'occurrent'</i> or RO relations such <i>'precedes'</i> or <i>'is about'</i>
primary property	Denotes properties that will appear grouped in the first block of properties in the data tool and in the search result page	service restrictions and fees, resource description.
embedded class	Denotes a class for which instances can only be created in the context of an embedding class	<i>'antibody immunogen'</i> created within <i>'antibody'</i> , <i>'construct insert'</i> created within <i>'plasmid'</i>
related lab	Denotes the properties that relate a resource to a laboratory	<i>'service provided by', 'located in'</i>
admin data	Denotes classes or properties that are never displayed in the search results	personal email, facilities address, last name and first name of facility contact persons

Table 1. Sample values for *inPropertyGroup* and *inClassGroup* properties.

Property Label	Description	Example	Property Type
eagle-i preferred label	Defines the value of preferred label to display in the data collection tool and search UIs	Capitalized 'Organization' for OBI_0000245 (<i>'organization'</i>)	Annotation Property
eagle-i preferred definition	Defines the value of preferred definition to be displayed in the data collection tool and search UIs	For OBI_0000245 (<i>'organization'</i>): "An entity that can play roles, has participants, and has a set of organizational rules."	Annotation Property
eagle-i domain constraint	Used to specify the domain of an imported property. Each annotation will contain the URI of one class.	Value set to "OBI_0000245" (<i>'organization'</i>) for RO property <i>'location_of'</i>	Data Property
eagle-i range constraint	Used to specify the range of an imported property. Each annotation will contain the URI of one class.	Value set to "ERO_0000004" (<i>'instrument'</i>) for RO property <i>'located_in'</i>	Data Property

Table 2. Additional properties defined in the UI Annotation Definition file.

The UI Annotations file has also been used to import external referenced classes that are used to populate drop-down menus in the data collection tool, such as MeSH terms for diseases. This file also contains shortcut relations between classes that in the core ontology are expressed using a more complex concatenation of properties to maintain full logical computability. For example, from an application standpoint we need to have a single

property that relates a service to a core laboratory providing that service. OBI uses a composed relation built from two properties to make this association between an organization and a service it provides (*'organization'* *'bearer_of'* some *'service provider role'* and *'realized_by'* some *'service'*). The UI Annotations file replaces this complex statement with a single property linking a service to its provider (*'service provider'*

'provides_service' some *'service'*) where *'service provider'* is defined as follows: [(*'organization'* or *'Homo sapiens'*) and (*'bearer_of'* some *'service provider role'*)]. This need to simplify complex relation chains will be a common issue in using ontologies for data collection applications, and approaches like the ones suggested in [12] should be exploited.

4 Discussion

In this section we discuss implementation choices and lessons learned from our effort to build an ontology that fulfills standards for interoperability and reuse and also requirements imposed by the UI and logic for the applications. The application ontology developed enables the annotation of eagle-i resources at the level of granularity and semantic complexity required for answering the queries in our use cases.

4.1 Use of Best Practices for Biomedical Ontologies Development

Upper ontology. We used BFO in our implementation. Use of an upper ontology can facilitate the design and modeling process, and ease the reuse of existent ontologies based on the same upper ontology. Issues can arise, however, when driving application interfaces directly from BFO-constructed ontologies. For example, our end users don't want to see BFO classes such as *'continuant'*, *'occurrent'* or *'processual entity'* in the UI. Accordingly, we used a *'Data Model Exclude'* annotation property to mask these BFO classes from appearing in the UI (See Table 1). While simple to implement, this solution requires the additional effort to create and maintain the annotations and to implement procedures and tools that can use them programmatically.

Another challenge was the fact that the domain and/or range of most RO properties are BFO classes. For instance, the RO property *'located_in'* has *'continuant'* as both its domain and range. We wanted to reuse RO properties, and at the same time present users with application-specific choices when filling values of properties in the eagle-i data collection tool, without forcing them to traverse all the subtrees under *'continuant'*. To achieve this, we used annotation properties to "restrict" the domain and range of those properties for use

within the data collection tool (see *'eagle-i domain restriction'* and *'eagle-i range restriction'* examples in Table 2). Finally, because we used properties to define the data fields that are shown for each resource type in the data collection tool UI, we had to exclude from the model those properties that were inherited from upper ontology classes. For example, *'transformation_of'* is inherited by all subclasses of *'continuant'*, but it was not appropriate to display this field in our UI for most continuants, such as *'laboratory'* or *'instrument'*.

Reusing existing ontologies and vocabularies. As described in Section 3 we have extensively used the MIREOT principle to reuse terms from external ontologies. One of the drawbacks we have found in applying MIREOT is related to the overhead of implementation efforts (creating, maintaining and running scripts to synchronize terms) and the lack of well-defined standards for custom axioms imports. Tools like OntoFox [13] or OWL Module Extractor² are helpful, but an ideal solution would be to integrate these tools within an ontology editor such as Protégé³.

Due to its wide usage for indexing related publications, we opted to import portions of MeSH to reference diseases. Development of best practices and better availability of standardized URIs for commonly used non-ontology based controlled vocabularies would enable better interoperability and reuse.

4.2 Layered eagle-i Ontology Modules

Our approach of decoupling application-specific from general purpose content through layered ontology modules has proved an effective means to drive an application UI while maintaining interoperability with external ontologies and data sources. Logically we wanted to separate our core ontology from the application-specific ontologies and therefore identify what was relevant to share with the community from what was specific to the needs of eagle-i. These layered modules also facilitated parallel development in a shared repository, as ontologists familiar with OWL constructs and functionality could manage

² <http://owl.cs.manchester.ac.uk/modularity/>

³ protege.stanford.edu

eagle-i core development, while curators were able to concurrently add proper annotation values in the UI annotations file.

We have identified a set of requirements for designing modular ontologies that can bridge the gap between an application and domain-specific ontologies. These include: (a) application-specific labels and definitions; (b) exclusion of sets of classes and properties from the model used by the application; (c) restriction of domain and range for some

imported properties; (d) definition of display order of object and data properties at class level. Figure 4 illustrates several of these features in the context of the eagle-i data annotation tool. Despite the effectiveness of this approach, it requires significant effort to keep the annotations current when the core module changes, and presents the risk of excessive proliferation of annotation properties and their instance values in attempts to simplify application coding complexity.

Figure 4. The eagle-i data collection tool user interface.

Shown is an example of a ‘*plasmid*’ record annotated using the eagle-i ontology. (1) eagle-i classes annotated with the “resource root” value are displayed in the left bar menu. (2) The value of ‘*eagle-i preferred definition*’ is used for tooltips that appear while hovering over the property labels. (3) The ‘*eagle-i preferred label*’ is used for the display name of property. Here, the imported RO ‘*location_of*’ has been renamed ‘*Location*’. This property is also flagged as a primary property using the ‘*inPropertyGroup*’ annotation property, as are ‘*Additional Name*’, ‘*Description*’ and ‘*Contact Person*’ properties. This flag results in presentation at the top of the property list for a record. (4) Users can select a technique associated with the reagent. In the ontology, the ‘*technique*’ class is annotated as a ‘*referenced class*’ which tells the UI to allow reference to an ontology term but create no instances. (5) Construct insert is an example of a resource annotated as an ‘*embedded class*’, which has to be created in the context of a construct or plasmid of which they are a part.

4.3 Coordination with the Biomedical Ontology Community

One of the most interesting aspects of the eagle-i “experiment” has been our commitment to collaboration with similar efforts aimed at resource modeling, data collection, and ontology development. We aligned our ontology with the NIF, VIVO, and RDS by reusing common terms and definitions, with the goal of migrating to usage of the same ontology classes and URIs at a later date. In doing so we have contributed to the design and enriched the content of existing biomedical ontologies – most notably OBI, where we contributed design patterns for services and added classes for devices, functions, and techniques. Finally, we developed ontological models for several important domains where the available ontologies are insufficient or non-existent, such as reagents, biospecimens, and genotype representation.

While the biomedical ontology community may benefit from these collaborations, it is clear that coordinated development practices require additional work for those involved in building or using ontologies to drive applications. It is not always straightforward to implement a consistent design within an application ontology that aligns and interoperates with external reference ontologies. For instance, eagle-i consulted with the OBI and NIF developer communities when modeling research-related services to ensure use of common principles and design patterns. However, the model that resulted, which classified services based on their input and output (e.g. data vs. a material entity), was not suitable for eagle-i users who preferred a hierarchy classified according to the process performed by the service (i.e. analysis, production, storage, etc). To accommodate both the broader biomedical ontology community and the eagle-i users, we implemented one service hierarchy in OBI, and then use MIREOT to reference individual service classes back into eagle-i and restructured them into a hierarchy that suited the needs of our application end-users.

It is preferable to maintain orthogonality of ontologies by having a single “home” for a particular kind of entity (like device in OBI or chemical reagent in ChEBI) [4]. However, this goal requires substantial dedication because

one first models in the application ontology where implementation can meet project requirements immediately. Next, term requests and/or modeling within an appropriate (reference) ontology are made, but this often takes a significant amount of time to coordinate agreement. Finally, the original model is obsoleted in the application ontology and classes from the reference ontology are referred using MIREOT. Additional effort is also required to keep terms synchronized.

The process of developing the eagle-i ontology modules has highlighted the need for effective automated mechanisms for extracting customized subsets of terms from reference ontologies. Such ‘customizable community views’ or ‘slims’ can be tailored to meet the needs of diverse communities or applications, and can minimize the effort involved in reusing existent ontologies [14]. Customized views can be extracted with different level of complexity (from referencing single terms through MIREOT, to importing a ‘refactored’ class hierarchy for a particular branch, or set of axioms to guarantee reasoning capabilities for a particular subset selected). This mechanism could also be used within an application ontology to provide a similar but more flexible implementation of our layered approach.

5 Conclusions

The process of developing an ontology-driven application has been an important benchmark for usage of biomedical ontologies and their driving design principles and tools. We have designed a layered set of modular ontologies, consisting of a broadly applicable core ontology and an application-specific ontology. This has allowed the identification of requirements and principles to inform a general design pattern for building applications that rely on ontologies for their logic and user interface.

We have identified a clear need to develop mechanisms, best practices and tools to bridge the gap between reference and application ontology development and usage. Future efforts will be aimed at refining and documenting these requirements, sharing our lessons learned, and engaging in efforts addressing the issues that the current biomedical ontology community is facing when dealing with real-world applications.

Acknowledgments

We acknowledge Ted Bashor, Rob Frost, Larry Stone and Daniela Bourges for their contributions to development of the eagle-i system.

References

1. Bug, W.J., et al.: The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, 6 (3), 175-94 (2008)
2. Tenenbaum JD. et al. The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. *J. Biomed. Inform.* Feb;44(1):137-45 (2011)
3. Torniai, C., Bashor, T., Bourges-Waldegg, D., Corday, K., Frost, H.R., Johnson, T., Segerdell, E., Shaffer, C.J., Stone, L., Wilson, M.L., Haendel, M.A.: eagle-i: an ontology-driven framework for biomedical resource annotation and discovery. In: *Bio-Ontologies 2010: Semantic Applications in Life Sciences*, ISMB, Boston (2010)
4. Smith, B., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25 (11), 1251-5 (2007)
5. Peters, B. and The OBI Consortium: Ontology for Biomedical Investigations. In: *International Conference on Biomedical Ontology*, Buffalo, 2009
6. Krafft, D.B. et al.: VIVO: Enabling National Networking of Scientists. In *Proceedings of the WebSci10*, Raleigh, NC (2010)
7. Horrocks, I., Patel-Schneider, P.F., van Harmelen, F.: From SHIQ and RDF to OWL: The making of a web ontology language. In: *Journal of Web Semantics* 1, 7-26 (2003)
8. Grenon, P., Smith, B.: SNAP and SPAN: Towards Dynamic Spatial Ontology. In: *Spatial Cognition & Computation: An Interdisciplinary Journal* 4, 69-104 (2004)
9. Ruttenberg, A.: Information Artifact Ontology (IAO), <http://code.google.com/p/information-artifact-ontology/>
10. Smith, B. et al.: Relations in biomedical ontologies. In: *Genome Biol* 6, R46 (2005)
11. Courtot, M., Gibson, F., and Lister, A.: MIREOT: the Minimum Information to Reference an External Ontology Term. *ICBO*, Buffalo (2009)
12. Mungall CJ, Ruttenberg A, Osumi-Sutherland D.: Taking shortcuts with OWL using safe macros. *Nature Precedings*. 2010
13. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y.: OntoFox: web-based support for ontology reuse. In: *BMC Research Notes*, 3:175 (2010)
14. Davis MJ, Sehgal MS, Ragan MA.: Automatic, context-specific generation of Gene Ontology slims. In: *BMC Bioinformatics*. 11:498. (2010)

Mapping Composition for Matching Large Life Science Ontologies

Anika Groß^{1,2}, Michael Hartung^{1,2}, Toralf Kirsten^{2,3}, Erhard Rahm^{1,2}

¹Department of Computer Science, University of Leipzig, Germany

²Interdisciplinary Centre for Bioinformatics, University of Leipzig, Germany

³Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Germany

Abstract. There is an increasing need to interrelate different life science ontologies in order to facilitate data integration or semantic data analysis. Ontology matching aims at a largely automatic generation of mappings between ontologies mostly by calculating the linguistic and structural similarity of their concepts. In this paper we investigate an indirect computation of ontology mappings that composes and thus reuses previously determined ontology mappings that involve intermediate ontologies. The composition approach promises a fast computation of new mappings with reduced manual effort. Our evaluation for large anatomy ontologies shows that composing mappings via intermediate hub ontologies is not only highly efficient but can also achieve better match quality than with a direct matching of ontologies.

Keywords: ontology matching, matching quality, compose, reuse

1 Introduction

Ontologies have become increasingly important for life sciences, in particular to semantically annotate molecular-biological objects such as proteins or pathways [5, 16]. There are frequently multiple interrelated ontologies of a domain. For instance, information about mammalian anatomy can be found in Foundational Model of Anatomy (FMA) [8], NCI Thesaurus (NCIT) [19], or Adult Mouse Anatomy (MA) [1]. This situation led to a growing need to determine mappings between pairs of related ontologies. These *ontology mappings* are valuable for enhanced semantic data analysis, data integration [11], for merging (combining) the ontologies [14] and to support comparative science, e.g., mouse models for human cancer [18].

Since the manual creation of mappings is often too labor-intensive for large ontologies with thousands of concepts, matching approaches have been proposed to (semi-) automatically determine ontology mappings, e.g., by calculating the linguistic or structural similarity between ontology concepts [7]. Many ontology matching systems have been developed in recent years and several of them

participate in the annual Ontology Evaluation Alignment Initiative (OAEL) [21]. A promising direction to determine ontology mappings that has found only little attention so far (see Related Work) is the reuse of previously determined mappings that involve the ontologies to be matched. In particular, the composition of mappings with an intermediate ontology can be used to indirectly compute ontology mappings. For instance, an ontology mapping between MA and NCIT can be obtained by composing two existing mappings to an intermediate ontology, e.g., the Uber anatomy ontology (Uberon) [25] or Universal Medical Language System (UMLS) [26]. Fig. 1 exemplifies this approach for two selected concepts *MA_0001421* and *NCI_C32239* that are described by name and synonym attributes. A direct match of both concepts is non-trivial since their names differ significantly. However, using Uberon as an intermediate ontology allows us to reuse the correspondences (matches) *MA_0001421-UBERON:0001092* (exact match of ‘*atlas*’) and *UBERON:0001092-NCI_C32239* (exact match of ‘*C1 vertebra*’). The composition of these two correspondences results in the new correspondence between *MA_0001421* and *NCI_C32239*.

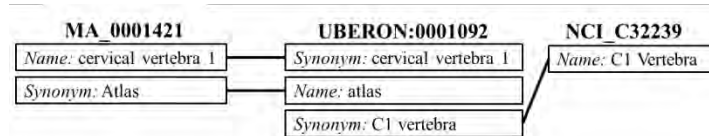


Figure 1. Composition of correspondences via an Uberon concept

Composing existing mappings for aligning two ontologies is promising in multiple ways. First, the life science community is continuously generating new mappings that are collected in repositories such as BioPortal [20] and can be reused for composition. Such a reuse is especially promising if the mappings are of high quality, e.g. validated by domain experts. Second, composing ontology mappings can likely be executed very fast while directly matching large ontologies is time-consuming and typically of quadratic complexity (every concept of the first is compared with every concept of the second ontology). Third, the intermediate ontology may contain additional knowledge which can be useful to detect further correspondences for improved mapping quality. Finally, if the intermediate ontology is a comprehensive ontology offering ontology mappings to numerous other ontologies we can consider it as a *hub ontology* providing large reuse potential. For a new ontology one only needs to create a mapping to the hub and use this for the composition to all other ontologies in the domain.

We therefore study composition-based matching of life science ontologies and make the following contributions:

- We propose a composition-based ontology matching approach which reuses previously determined mappings with one or multiple intermediate ontologies.
- Our approach is based on powerful ontology and mapping operators such as compose, match and extract. The approach also supports an incremental extension of composed mappings for improved match quality.
- We evaluate the approach by determining ontology mappings between the anatomy ontologies MA and NCIT utilizing large intermediate ontologies such as UMLS, FMA, Uberon and RadLex. The results demonstrate the high effectiveness and efficiency of composition-based ontology

matching.

In Sec. 2, we introduce our ontology/mapping model and discuss the ontology matching process. Sec. 3 defines the operators and shows their use within our composition-based match approach. We evaluate the approach in Sec. 4. After a discussion of related work (Sec. 5), we summarize and outline possible future work.

2 Preliminaries: Models and Ontology Matching

An ontology $O = (C, R)$ consists of a set of concepts C that are interconnected by relationships $r \in R$. Each ontology concept is described by a set of single- or multi-valued attributes. The concept name is the most common attribute in life science ontologies. Further common attributes include synonym (alternate name) and concept definition. A special attribute accession number c_{acc} is used to unambiguously identify a concept c within the ontology O . Relationships interconnecting concepts are of a certain kind such as *is_a* (e.g., ‘lung’ *is_a* ‘organ’) to represent specialization relationships or *part_of* (e.g., ‘left lung’ *part_of* ‘lung’) to represent part-whole relationships.

An ontology mapping $M_{O_1, O_2} = \{(c_1, c_2, sim) \mid c_1 \in O_1, c_2 \in O_2, sim \in [0, 1]\}$ between two ontologies O_1 and O_2 consists of a set of correspondences. Each correspondence (c_1, c_2, sim) interconnects two related or equivalent ontology concepts c_1 and c_2 . The strength of the connection is represented by a similarity value *sim* between 0 and 1. The greater the similarity value, the more similar are the corresponding concepts.

Ontology mappings can be created manually by domain experts. However, the complexity and size of the input ontologies make a manual creation often impossible. Thus, (semi-)automatic ontology matching approaches have been proposed [7]. They can roughly be classified into metadata- or schema-

based and instance-based approaches [23]. The metadata-based approaches are mostly either linguistic or structural matchers. Linguistic matchers typically employ string similarity measures, e.g., ExactMatch, n-gram or EditDistance, on the concept attributes (concept name, synonym, definition). Structural matchers also consider the ontology structure for matching, e.g., context information from children or ancestors of concepts. To improve the match quality compared to the adoption of a single matcher, current match systems such as COMA++ [3] or GOMMA [10, 13] support the flexible combination of multiple matchers and the aggregation of their results. In this paper we focus on linguistic matchers since previous works [9] has shown that they produce ontology mappings of good quality especially for anatomy ontologies that we consider here.

3 Mapping Composition

In this section, we present our composition-based match algorithm to indirectly match ontologies by reusing existing ontology mappings. We start with a discussion of the general idea of using intermediate ontologies in Sec. 3.1 and introduce our ontology and mapping operators in Sec. 3.2. We then combine the proposed operators in the composition-based match algorithm (Sec. 3.3).

3.1 Indirect Matching Via Intermediate Ontologies

The general idea of our approach is to use mappings to intermediate ontologies for indirect ontology matching. Such mappings

are typically produced in a resource-intensive match process, in particular, when the mappings or portions of them are created manually or computed by sophisticated match algorithms. Therefore, reusing such mappings promises to save or reduce the huge effort necessary when starting from scratch for matching two ontologies. Repositories such as BioPortal provide an increasing number of ontology mappings that can be used for a composition-based ontology matching.

Fig. 2a shows the basic situation consisting of two ontologies O_1 and O_2 as well as mappings from O_1/O_2 to several intermediate ontologies IO_1, \dots, IO_k . Intermediate ontologies should have a significant overlap with the ontologies to be matched, i.e. the mappings should contain correspondences for a larger part of the ontologies' concepts. If possible it is reasonable to use the knowledge from different intermediate ontologies as they may complement each other. As composition of ontologies is likely very fast it is easily feasible to determine composed mappings for several intermediate ontologies.

In some cases (Fig. 2b), there is a centralized hub ontology HO that is predominant in the domain. Typically, such an ontology has many mappings to other ontologies. Any new ontology O_{new} can then be aligned with any other ontology O_1, \dots, O_n by first matching O_{new} to HO . Afterwards, the mapping $M_{O_{new},HO}$ can be composed with any available mapping M_{HO,O_i} in the domain. Hence, aligning the ontology O_{new} with any ontology O_i can be efficiently computed.

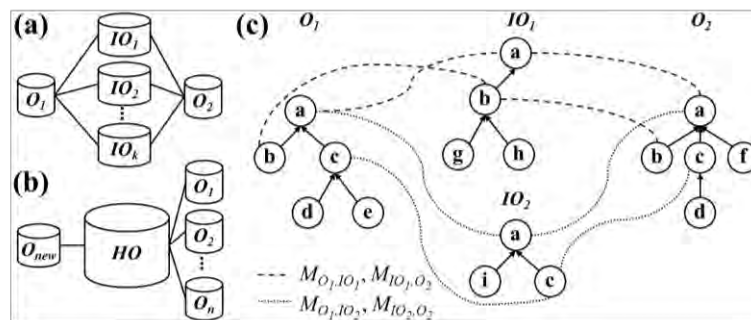


Figure 2. Mapping composition via intermediate ontologies IO_1, \dots, IO_n (a)
Match new ontology to hub ontology (b)
Example for composition-based ontology matching (c)

3.2 Operators

Previous research on the generic management of models and mappings [4] has already identified a series of operators that can be adapted for our purpose of ontology matching. In the following, we introduce the ontology and mapping operators match, compose and extract. Furthermore, we use merge to combine several (composed) mappings.

The match operator matches the concepts of an ontology O_A against the concepts of a second ontology O_B and directly determines an ontology mapping M_{AB} consisting of correspondences with similarity values (sim) between 0 and 1.

$$\begin{aligned} \text{match}(O_A, O_B): O_A \times O_B &\rightarrow M_{AB} \\ M_{AB} = \{(c_1, c_2, sim) \mid c_1 \in O_A, c_2 \in O_B\} \end{aligned}$$

The compose operator allows for the composition of mappings, i.e., it combines two mappings (M_{AB}, M_{BC}) to indirectly determine a new mapping (M_{AC}). Two correspondences of different mappings can be composed to a new correspondence if the range concept of the first correspondence is equal to the domain concept of the second correspondence. Different functions can be used to aggregate the similarity values of correspondences ($aggSim$), e.g., by computing the average or maximum similarity.

$$\begin{aligned} \text{compose}(M_{AB}, M_{BC}): M_{AB} \times M_{BC} &\rightarrow M_{AC} \\ M_{AC} = \{(c_1, c_2, aggSim(sim_1, sim_2)) \mid c_1 \in O_A, b \in O_B, \\ c_2 \in O_C: \exists (c_1, b, sim_1) \in M_{AB} \wedge \exists (b, c_2, sim_2) \in M_{BC}\} \end{aligned}$$

The extract operator reduces an ontology O_A to a delta ontology ΔO_A by returning only those concepts that are not covered by an input mapping M_{AB} between O_A and another ontology O_B . It can be used to match only the delta ontologies ($\text{match}(\Delta O_A, \Delta O_B)$) to save the comparisons that are already covered by the (partial) mapping M_{AB} .

$$\begin{aligned} \text{extract}(O_A, M_{AB}): O_A \times M_{AB} &\rightarrow \Delta O_A \\ \Delta O_A = \{c \mid c \in O_A, \nexists b \in O_B: (c, b, sim) \in M_{AB}\} \end{aligned}$$

The merge operator aggregates several input mappings between the same ontologies to a combined mapping. The merge decision is based on a minimum occurrence count occ in the k input mappings, i.e., a correspondence must appear in at least occ of the k input mappings. Note that $occ=1$ corresponds to a standard union whereas $occ=k$ corresponds to

the intersection of all mappings.

$$\begin{aligned} \text{merge}(M_{AB1}, \dots, M_{ABk}, occ): M_{AB1} \times \dots \times M_{ABk} \times occ &\rightarrow \\ M_{AB} \\ M_{AB} = \{(c_1, c_2, aggSim) \mid (c_1, c_2, sim) \text{ occurs in at least} \\ occ \text{ mappings of } M_{AB1}, \dots, M_{ABk}\} \end{aligned}$$

3.3 Composition-Based Match Approach

The introduced operators are used within two algorithms that make up our composition-based match approach: *composeMatch* and *extendMatch*. *composeMatch* takes as input two ontologies O_1 and O_2 , a list of one or more intermediate ontologies IO_1, \dots, IO_k as well as the parameter occ denoting the occurrence count for mapping merge. The algorithm produces a composed mapping between O_1 and O_2 using the intermediate ontologies by reusing existing mappings. Firstly, for each intermediate ontology IO_i in the list we get the mappings between O_1 and IO_i as well as between IO_i and O_2 , e.g., from a repository. Afterwards the compose operator is applied to the mappings $M_{O1, IO_i}, M_{IO_i, O2}$ to determine a mapping between O_1 and O_2 . This composed mapping is added to the list of mappings (*MapList*). Finally, all mappings in *MapList* are merged to a combined mapping controlled by parameter occ . The merge of several mappings likely improves match quality. For example, the union of complementing intermediate ontologies can help to find more correct correspondences thereby improving recall. If the input list contains only one intermediate ontology, e.g., a known hub, the merge step can be omitted.

Typically, a composed mapping $CM_{O1, O2}$ may not cover all parts of the ontologies O_1 and O_2 that need to be matched. Therefore, the algorithm *extendMatch* can be applied optionally to further improve recall and match quality. It takes the two ontologies as well as the composed mapping as input. To find additional correspondences between unmatched ontology parts we use the extract operator to determine the sub-ontologies of O_1 and O_2 that are not covered by $CM_{O1, O2}$. The resulting delta ontologies $\Delta O_1, \Delta O_2$ are matched directly using a specific match algorithm, e.g., string similarity of the attributes name and synonym. We then determine the union (merge with $occ=1$) of this direct mapping $DM_{\Delta O1 \Delta O2}$ and the composed mapping $CM_{O1, O2}$. Note that, all produced mappings can be filtered by selection

Algorithm *composeMatch*($O_1, O_2, IO_1 \dots IO_k, occ$)

Input: Two ontologies O_1 and O_2 , list of intermediate ontologies $IO_1 \dots IO_k$, occurrence count occ

Output: Composed mapping CM_{O_1, O_2}

```
1: MapList  $\leftarrow$  empty
2: for each  $IO_i \in IO$  do
3:    $M_{O_1, IO_i} \leftarrow \text{getMapping}(O_1, IO_i)$ 
4:    $M_{IO_i, O_2} \leftarrow \text{getMapping}(IO_i, O_2)$ 
5:   MapList.add(compose( $M_{O_1, IO_i}, M_{IO_i, O_2}$ ))
6: end for
7: return merge(MapList,  $occ$ )
```

criteria (e.g., a minimal similarity threshold) or advanced post-processing steps to improve precision, i.e., to reduce the number of incorrect correspondences.

Figure 2c illustrates an exemplary application of *composeMatch* for matching ontologies O_1 and O_2 via two intermediate ontologies IO_1 and IO_2 . Dotted lines represent the correspondences of O_1 and O_2 to the intermediate ontologies. The mapping composition (line 2-6) will produce the following *MapList* with two mappings between O_1 and O_2 consisting of two correspondences each: $\{(a,a), (b,b)\}, \{(c,c), (a,a)\}$. The merge aggregation of the *MapList* (line 7) with $occ=1$ results in the union mapping $\{(a,a), (b,b), (c,c)\}$ whereas $occ=2$ leaves only a single correspondence $\{(a,a)\}$. Not shown are the similarity values that need to be aggregated, e.g., by computing the average similarity.

4 Evaluation

4.1 Evaluation Setup

In all experiments, we align the Adult Mouse Anatomy ontology (MA) with the anatomical part of the NCI Thesaurus (NCIT). This match task is part of the annual OAEI contest so that the perfect mapping can be used for evaluating the quality (precision, recall and its combination F-measure) of the generated mappings. Mapping composition is performed with the help of four large intermediate ontologies, namely FMA [8], Uberon [25], RadLex [22], and UMLS [26] in their versions of late 2010. Table 1 summarizes statistical properties of the utilized ontologies and

Algorithm *extendMatch*(O_1, O_2, CM_{O_1, O_2})

Input: Two ontologies O_1 and O_2 , composed mapping CM_{O_1, O_2}

Output: Extended Mapping EM_{O_1, O_2}

```
1:  $\Delta O_1 \leftarrow \text{extract}(O_1, CM_{O_1, O_2})$ 
2:  $\Delta O_2 \leftarrow \text{extract}(O_2, \text{inverse}(CM_{O_1, O_2}))$ 
3:  $DM_{\Delta O_1 \Delta O_2} \leftarrow \text{match}(\Delta O_1, \Delta O_2)$ 
4:  $EM_{O_1, O_2} \leftarrow \text{merge}(\{CM_{O_1, O_2}, DM_{\Delta O_1 \Delta O_2}\}, 1)$ 
5: return  $EM_{O_1, O_2}$ 
```

mappings. The ontologies significantly differ in their total number of concepts ($|C|$) and the number of name/synonym attributes per concept ($\emptyset\text{NameSyn}$) (Table 1a). All intermediate ontologies are significantly larger than MA and NCIT. The ontology mappings used for the algorithm *composeMatch* have been computed once based on the linguistic similarity (trigram with threshold 0.8) of concept names and synonyms. Hence we compose automatically determined mappings instead of manually verified ones making it more difficult to achieve high mapping quality. Table 1b reveals significant differences in the mapping coverage (Cov) and sizes ($|Map|$) for MA and NCIT. For UMLS and Uberon, the mappings cover up to 80% and more while RadLex is limited to about 40%, i.e. this intermediate ontology cannot provide correspondences for most concepts. The FMA mappings have only medium coverage potentially influenced by relatively few names and synonyms per concept (Table 1a) limiting the quality of linguistic matching. By contrast, Uberon is a promising intermediate ontology due to its high $\emptyset\text{NameSyn}$ value. We generally expect ontologies providing many synonym terms to be adequate intermediate ontologies w.r.t. linguistic matching.

Applying *extendMatch* determines the two Δ -ontologies: $\Delta O_1 = \{d, e\}$, $\Delta O_2 = \{d, f\}$ (line 1-2) which are then matched against each other (line 3). The resulting direct mapping, e.g., $DM_{\Delta O_1 \Delta O_2} = \{(d,d)\}$, is merged with the input mapping CM_{O_1, O_2} so that the final mapping $\{(a,a), (b,b), (c,c), (d,d)\}$ is obtained.

(a)

	C	Ø NameSyn
MA	2,738	1.1
NCIT	3,298	2.5
Ueberon	4,958	4.9
UMLS	87,913	3.1
RadLex	30,773	1.6
FMA	81,059	1.8

(b)

Mapping	Cov _{Domain}	Cov _{Range}	Map
MA-Ueberon	80%	45%	2300
Ueberon-NCIT	33%	48%	1703
MA-UMLS	69%	3%	2975
UMLS-NCIT	5%	87%	4214
MA-RadLex	39%	3%	1082
RadLex-NCIT	4%	40%	1347
MA-FMA	57%	2%	1601
FMA-NCIT	3%	67%	2337

Table 1. Statistics for ontologies (a) and mappings (b)

The match operation within the *extendMatch* algorithm and the direct match computation consists of the steps pre-processing, similarity calculation, and post-processing. Pre-processing includes the elimination of delimiters and stop words, transformation to lower case letters, and word stemming. The similarity between ontology concept pairs is calculated based on the linguistic trigram similarity on concept names and synonyms. Post-processing consists of a MaxDelta selection [6] of correspondences returning for a concept the correspondences with the maximal similarity value or within a small delta distance to the maximal value. Furthermore, correspondences must meet a so-called CrissCross condition [12] for improved precision that eliminates conflicting correspondences (a_1, b_1) and (a_2, b_2) where a_2 is a child of a_1 but b_1 a child of b_2 or vice versa.

4.2 Composition-based matching

We first compare the quality of indirectly determined ontology mappings using the *composeMatch* algorithm as well as the impact of *extendMatch*. Fig. 3 summarizes the obtained mapping quality in terms of F-measure and compares them with the quality of a direct match (called as *no IO*). The direct matching based on linguistic similarity achieved a F-

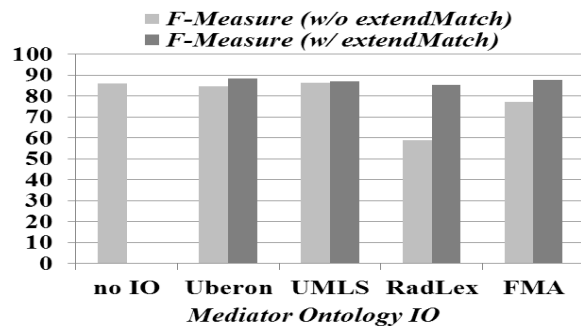


Figure 3. Compose via different intermediate ontologies

measure of 86% which is comparable to the best value of all previous OAEI contests (87.7%). The quality of the composed mappings (light grey bars) strongly depends on the utilized intermediate ontology and their associated mappings. The best F-measure values are achieved for composition via UMLS (86.2%) and Uberon (84.7%). Particularly, the UMLS-based mapping even exceeds the quality achieved by a direct match. Ontology mappings using FMA (77%) and RadLex (59%) only achieve a low quality influenced by the low coverage of their mappings to MA and NCIT. While RadLex is not primarily concerned with anatomy, Uberon provides a cross-species anatomy ontology and UMLS contains a huge anatomy part making these ontologies highly suitable for indirectly matching anatomy ontologies. The dark grey bars in Fig. 3 denote the achieved quality by an additional application of *extendMatch*. The results indicate that this additional step always leads to an improved quality. Interestingly, Uberon now achieves the best quality (88.2%) exceeding UMLS (87.0%) and the best OAEI result so far. This indicates that composition via Uberon finds non-trivial correspondences that cannot be identified by a direct match. The additional match effort of *extendMatch* improves match quality especially for intermediate ontologies with comparatively poor compose results (e.g., RadLex and FMA).

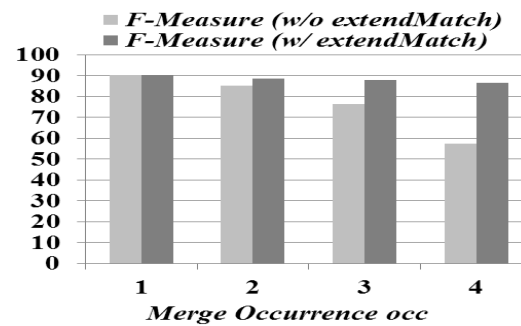


Figure 4. Merge composed mappings

Next, we determine whether the combination of several composed mappings from different intermediate ontologies improves match quality. Figure 4 shows the resulting F-measure values when merging the four composed mappings for different values for the occurrence count *occ* (specifying how often a correspondence has to occur in the individual composed mappings). The results show that merging several composed mappings improves match quality to up to 90.2% F-measure (recall 87.8%; precision 92.7%) for *occ*=1, i.e. when we take the union of the mappings. So we can outperform the quality of direct matching as well as the best OAEI result by our composition-based approach although we only compose automatically determined mappings. The intersection of the mappings (*occ*=4) turned out to be not effective (F-measure 57.4%) due to a significant reduction of recall, i.e. we can no longer take advantage of complementary correspondences provided by different intermediate ontologies. Additionally applying *extendMatch* leaves the result for *occ*=1 almost unchanged (90.3%) while it can significantly improve match qualities for larger *occ* values. Hence, the union of composed mappings can be used without applying an extra matching step. None of the previous approaches participating in OAEI anatomy track could exceed 87% F-Measure such that an increase to more than 90% is a significant improvement.

The execution times of the match process (without parsing ontologies/mappings) could be significantly reduced. The compose via all four intermediate ontologies and the following mapping merge took only 2.8s. The execution time for the additional *extendMatch* was 36s on average. By contrast the full direct match of the whole ontologies took 116s, i.e., the execution time could be reduced by up to a factor of 41 while achieving similar or even better match quality.

5 Related Work

The direct matching of large life science ontologies has been studied before [9, 15, 21]. Thereby different match approaches such as lexical and structural methods have been evaluated, e.g., in the domain of anatomy [28, 17].

The operators compose, match and extract

were introduced within the framework of model management [4]. They can be used in scenarios such as schema evolution to adapt dependent artifacts like instance data and views. In contrast we use these operators to efficiently match two ontologies based on composition.

The match compose operation has been introduced in schema matching before [3, 6] but was not applied for ontology matching. So far, there has been some attention on indirect matching and mapping composition in the life sciences. [27] derived indirect mappings using FMA as reference ontology. By contrast, we focus on using multiple complementing intermediate ontologies as well as an additional *extendMatch* to improve recall of compose. [24] presented an empirical analysis of mapping composition. They analyzed a pool of mappings without distinguishing different intermediate ontologies. Hence, it was not the focus to study which ontologies are useful hub ontologies. [2, 15] matched ontologies or other vocabularies by using a single ontology as domain/background knowledge. These strategies differ from our approach as they do not combine the knowledge of several different intermediate ontologies.

6 Conclusion and Future Work

We proposed a composition-based approach for indirectly matching life science ontologies via one or several intermediate ontologies. The goal is to reuse previously determined ontology mappings for improved match efficiency and quality. The approach is based on ontology and mapping operators compose, match, extract and merge. It allows the flexible combination of several composed mappings and the incremental extension of mappings by additional match steps for unmatched ontology concepts.

In our evaluation for large anatomy ontologies we considered four intermediate ontologies, namely UMLS, FMA, Uberon and RadLex. Overall, we achieved very good match quality (>90%) and significantly reduced execution times using a composition-based match instead of a direct match strategy. While the use of *extendMatch* is generally helpful to improve match quality, mapping composition alone was able to outperform the runtime and quality compared to direct matching especially when we merge several composed mappings.

Uberon and UMLS showed to be very effective intermediate ontologies and are thus suited as hub ontologies in the anatomy domain.

In future work, we plan to investigate composition-based ontology matching for further domains. We also want to study the impact of considering additional mappings, e.g. determined by structural matching or existing mappings from BioPortal. Furthermore, we want to investigate when it could be useful to compose more than two mappings within longer mapping chains.

Acknowledgments

This work is supported by the German Research Foundation (DFG), grant RA 497/18-1 ("Evolution of Ontologies and Mappings").

References

1. Adult Mouse Anatomy: http://www.informatics.jax.org/searches/AMA_form
2. Aleksovski, Z., Klein, M., Ten Kate, W., Van Harmelen, F.: Matching unstructured vocabularies using a background ontology, Managing Knowledge in a World of Networks, 182-197, Springer (2006)
3. Aumüller, D., Do, H.H., Massmann, S., Rahm, E.: Schema and ontology matching with COMA++. Proc. SIGMOD (2005)
4. Bernstein, P.A., Melnik, S.: Model management 2.0: manipulating richer mappings. Proc. SIGMOD (2007)
5. Bodenreider, O., Stevens, R.: Bio-ontologies: current trends and future directions. Briefings in Bioinformatics 7(3), 256-274 (2006)
6. Do, H.H. and Rahm, E.: COMA: a system for flexible combination of schema matching approaches. Proc. VLDB (2002)
7. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer (2007)
8. Foundational Model of Anatomy: <http://sig.biostr.washington.edu/projects/fm/>
9. Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating mappings for ontologies in biomedicine: Simple methods work. Proc. AMIA Annual Symposium (2009)
10. Gross, A., Hartung, M., Kirsten, T., Rahm, E.: On Matching Large Life Science Ontologies in Parallel. Proc. Data Integration in the Life Sciences (2010)
11. Jakoniene, V., Lambrix, P.: Ontology-based integration for bioinformatics. Proc. ODBIS Workshop @ VLDB (2005)
12. Jean-Mary, Y.R., Shironoshita, E.P., Kabuka, M.R.: Ontology matching with semantic verification. Journal of Web Semantics 7(3), 235-251 (2009)
13. Kirsten, T., Gross, A., Hartung, M., Rahm, E.: GOMMA: A Component-based Infrastructure for managing and analyzing Life Science Ontologies and their Evolution, Submitted to Journal of Biomedical Semantics (2011)
14. Lambrix, P., Edberg, A.: Evaluation of ontology merging tools in bioinformatics. In: Proc. of the 8th Pacific Symposium on Biocomputing (2003)
15. Lambrix, P., Tan, H.: SAMBO - A system for aligning and merging biomedical ontologies. Journal of Web Semantics 4(3), 196-206 (2006)
16. Lambrix, P., Tan, H., Jakoniene, V., Strömbäck, L.: Biological Ontologies. Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences (2007)
17. Mork, P., Bernstein, P.A.: Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. Proc. ICDE (2004)
18. Mouse Models of Human Cancers Consortium: <http://www.nih.gov/science/models/mouse/resources/hcc.html>
19. NCI Thesaurus: <http://ncit.nci.nih.gov/>
20. Noy, N.F. et al.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37 (Web Server Issue), W170-W173 (2009)
21. Ontology Alignment Evaluation Initiative: <http://oei.ontologymatching.org>
22. RadLex: <http://www.radlex.org>
23. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal 10(4), 334-350 (2001)
24. Tordai, A. et al.: Lost in Translation? Empirical Analysis of Mapping Compositions for Large Ontologies. Proc. Ontology Matching Workshop @ ISWC (2010)
25. Uber Anatomy Ontology: http://obofoundry.org/wiki/index.php/UBERON:Main_Page
26. Unified Medical Language System: <http://www.nlm.nih.gov/research/umls>
27. Zhang, S., Bodenreider, O.: Alignment of multiple ontologies of anatomy: Deriving indirect mappings from direct mappings to a reference. AMIA Annual Symposium (2005)
28. Zhang, S., Bodenreider, O.: Experience in aligning anatomical ontologies. Semantic Web and information systems 3(2), 1-26 (2007)

Generic Semantic Relatedness Measure for Biomedical Ontologies

João D. Ferreira, Francisco M. Couto

Departamento de Informática Faculdade de Ciências da Universidade de Lisboa Campo Grande, Lisboa, Portugal

Abstract. This paper presents a new method to measure semantic relatedness between concepts of an ontology with a rich set of relationship types, and performs a preliminary assessment of its validity. The measure was designed to be applicable to all biomedical ontologies, and to be flexible enough as to allow for different applications to address their own requirements by tuning, for example, the weight of each relationship type. We focus on the fact that we measure relatedness instead of similarity, which measures not simple likeness between concepts but also other interactions like articulation of cartilage and bones. We applied the measure to the Foundational Model of Anatomy, an ontology of human anatomy, and showed that it can be used to differentiate between related pairs of anatomical concepts and unrelated ones with higher performance than a custom similarity measure would. This work has shown positive preliminary analysis of the generic measure developed, which is a step forward to implementing tools to process the information contained in the increasing amount of biomedical ontologies.

Keywords: Semantic similarity and relatedness, Biomedical ontology applications, Relationship types, FMA

1 Introduction

One of the most important techniques applied to biomedical ontologies has been the calculation of semantic similarity between concepts, which quantifies the similarity in meaning between two concepts, as described in the ontology [15]. This technique is at the core of many applications, such as information search [2], where the results are sorted from most similar to the original query to least similar. Without a framework that enables computers to grasp the concept of semantic similarity, it would be impossible to automatically understand that, e.g., “Heart” is more similar to “Kidney” than to “Toenail”.

Semantic similarity, however, does not serve all purposes; in some cases, relatedness measures provide a more interesting and effective way of solving a problem. For instance, in the study of localized diseases, the physical proximity between anatomical concepts can be more meaningful than their similarity; also, from the point of view of pharmacology, it is meaningful that both

“lisuride” and “metixene” (identifiers chebi:51164 and chebi:51024 respectively) are antiparkinson drugs, despite their lack of structural similarity.

Pedersen *et al.* [13] mention the difference between *similarity* and *relatedness*. According to them, similarity is a stricter form of relatedness: a pair of similar concepts share form, shape or structure: in other words, the concepts are *alike*. As such, similarity is strongly related to relationships of subsumption: “Heart” *is-a* “Organ”, just like “Kidney” *is-a* “Organ”, etc. On the contrary, this paper presents a semantic *relatedness* measure that takes into account various relationship types, particularly in ontologies that use several, such as in the biomedical domain [17]. For example, the Foundational Model of Anatomy (FMA), where single inheritance could potentially lead to smaller levels of expressiveness, contains over 60 relationship types, effectively allowing for a rich content and a high expressiveness; other examples include ChEBI, with 10, and PATO, with 8.

2 State of the Art

In the biomedical field, semantic similarity has been extensively used in the Gene Ontology (GO) [10, 15], with applications like prediction of protein function [11, 6], prediction of protein-protein interactions [19] or prediction of breast cancer outcome [18]. Other ontologies used with semantic similarity approaches in the biomedical domain include the HPO, where similarity of phenotypes is used to “refine the differential diagnosis by suggesting clinical features that, if present, best differentiate among the candidate diagnoses” [8], and ChEBI, where it was used to predict properties of small molecules [5]. Even though the methods described in these papers are named *similarity* measures, most of them use relationship types other than subsumption, which effectively makes them semantic *relatedness* measures.

However, relatedness measures, explicitly named so, have not been very well explored in the biomedical domain. Patwardhan et al. [12] explore a relatedness measure in WordNet, based on context vectors that represent co-occurrence of words, which was later adapted [13] to work with SNOMED-CT, a clinical terminological resource, but this method does not use the relations of the ontology, only the concepts themselves and their descriptors.

3 Generalization of Relatedness Measures

Semantic relatedness suffers from a lack of generalization, as the methods currently in use have all been specifically tailored to work on the ontologies they are applied to, and greatly depend on the subsumption of concepts. With the establishment of OBO and the increasing interest in ontology development and application to real-world problems, we are at a point where relatedness measures are expected to be developed to other ontologies as well. This problem can be handled in two distinct but parallel perspectives: we can wait for a team to develop, evaluate, publish and deploy a measure to work with their own ontologies, and/or we come forward with a generalized methodology that can be applied to all biomedical ontologies.

Specifically developing a measure for an ontology is, perhaps, the preferred solution: semantic relatedness measures tailored to one ontology will most likely deliver best performance than a general methodology. However, if a strong generalized measure is developed, information retrieval teams can build their systems over it without the need to create a specific measure from scratch for each ontology of interest. Furthermore, for research on epidemiologic surges, a field that uses ontologies of, e.g., diseases, symptoms, anatomical parts and geographic locations [3], readily applicable measures of semantic relatedness would be an asset for a quick deployment of results.

We present a measure of relatedness that can be easily applied to all biomedical ontologies, as long as they define concepts and relations between them. It is flexible enough to allow for a number of adaptations that can be fine tuned not only for the ontology itself but also according to the type of application making use of the measure. As a case study, we applied the measure to FMA, a complex ontology where the methods that have been used in GO do not work well (see the section on Results).

3.1 Relatedness Measure

In general, similarity can be calculated based on the *is-a* relationship. For example, “Heart” is more similar to “Kidney” (both are organs) than to “Cardiac ventricle”. Instead, to measure *relatedness*, we propose a metric that takes into account not the likeness of two concepts but the overlap of their neighborhoods.

The formula for the relatedness measure we propose depends on the relevance of one concept to another one. We use a relevance factor, $\omega(i \rightarrow x)$, to express the relevance of concept i with relation to concept x , and take $N(x)$ as the neighborhood of x (the concepts that are relevant to x). Relatedness between two concepts, $\rho(A, B)$ is then measured through the overlap in their neighborhood:

$$\rho(A, B) = \frac{\sum_{i \in N(A) \cap N(B)} \omega(i \rightarrow A) + \omega(i \rightarrow B)}{\sum_{i \in N(A) \cup N(B)} \omega(i \rightarrow A) + \omega(i \rightarrow B)} \quad (1)$$

with $\omega(i \rightarrow x) = 0$ if $i \notin N(x)$.

Equation 1 can be adapted to a wide

number of situations. For example, $N(x)$ can be defined as the set of concepts connected to x with a path of at most M relations (the radius of the neighborhood). $\omega(i \rightarrow x)$ can be defined based on the relationship types of the path from i to x : if this path is composed of n relations of type r_1, \dots, r_n , then

$$\omega(i \rightarrow x) = \prod \text{weight}(r_j) \quad (2)$$

where weight of the relations is higher for more important relationship types.

Other examples include $N(x)$ as the set of concepts whose relevance factor is above a certain threshold; or the relevance factors can be fine tuned to measure relevance taking into account specificity (e.g., through information content). In fact, if one takes $N(x)$ to be the set of superclasses of x and $\omega(y \rightarrow x)$ to be the information content of y , the measure is not very different from sim_{GIC} (the difference being that common superclasses would appear twice in the numerator and in the denominator) [14].

Each application is free to define what constitutes a neighborhood and what is the relevance of one concept to another one. For instance, in an application concerned more with physical location than with similarity, the following should hold:

$$\begin{aligned} \omega(\text{"Cardiac ventricle"} \rightarrow \text{"Heart"}) \\ > \omega(\text{"Kidney"} \rightarrow \text{"Heart"}) \end{aligned}$$

3.2 FMA

The Foundational Model of Anatomy (FMA)¹ [16] is restricted through single inheritance but its many relationship types make it a very content-rich ontology. As per the definition in [13], these relations can be exploited to determine the relatedness between two anatomical concepts. For instance, the relationship type *articulates-with*, which “holds between two or more adjacent bones or between a bone and a cartilage through a joint” (from the definition of FMAID:276393), does not convey likeness between the connected

concepts, but there is no doubt that concepts connected through it are related. Whether this relationship is important for an application is dependant on that application’s goal, and as such, any measure should be flexible enough to allow the user to determine which relationship types are relevant, and to what extent.

We chose FMA as our case study based on four points:

1. it uses over 60 relationship types, which means a lot of semantic information is contained in non-subsumption relations;
2. it is a very complete ontology of the human anatomy, with applications such as X-ray and disease annotation. In fact, we have used cross references between FMA concepts and diseases to assess the validity of the developed measure;
3. we plan to extend this measure to ontologies in the epidemiological field in the future, and FMA is one such ontology;
4. semantic similarity measures developed for GO do not deliver good results in this ontology.

In this case study, $N(x)$ was defined as the concepts that are connected to x through a path no longer than M relations (where $M \in \{3, 4\}$), and $\omega(i \rightarrow x)$ was defined through equation 2 with the weight of all relationship types set to 0.7. For concepts connected with more than one path, $\omega(i \rightarrow x)$ is the maximum of all those relevance factors.

On the one hand, a small value of M makes the intersection of neighborhoods likely to be empty. By increasing its value, the measure considers a larger neighborhood and gains resolution. On ontologies with more concepts or less relations, M should be increased to avoid that problem. On the other hand, a large value of M increases the time to calculate relatedness, so a compromise must be made. We studied values of $M \in \{3, 4\}$ to understand that compromise.

As for the weight of each relation, it was verified that changing the absolute value did not particularly influence the results. In a specific application, the relative weight of a relation must be attributed based on its relevance; as an example, we used a value of 0.7 since it decreases the relevance factor of

¹ Accessible from: <http://sig.biostr.washington.edu/projects/fm/AboutFM.html>. The ontology is described in frames and must be opened with the Protégé software or used directly as a MySQL database. OBO and OWL version exist, but they contain only a subset of the ontology.

the more distant concepts, but values of 0.5 and 0.8 showed the same kind of results.

4 Results and Discussion

Our measure returns a value of relatedness between two FMA concepts. In order to assess the validity of such a measure we must determine if the value returned makes sense in a biomedical context. As discussed above, each application must weight each relationship type depending on its main goal. For this first measure of assessment, however, we have assigned equal weights to all relationships.

Two avenues were pursued to validate the approach. The first was based on a simple match between FMA and GO. There is an overlap of 274 labels in both ontologies (counting preferred names and synonyms), with 256 GO cellular component concepts matched to 267 FMA concepts. Using these matches, we were able to compare FMA's relatedness measure with two of the most successful similarity measures developed for GO, Resnik and sim_{GIC} [14]. Figure 1 shows the scatter plots, where the X-axis has FMA's relatedness measure and the Y-axis has GO's similarity measure. Correlation coefficients are given for each plot in Table 1. There is some correlation between the two measures, which is a good indication of the validity of the proposed method. However, these values are only relatively good, given that:

1. first and foremost, we are comparing a similarity measure with a relatedness one;
2. the two ontologies take distinct points of view about the cellular domain [1];
3. sim_{GIC} and Resnik use external background knowledge (in the form of information content) and this measure uses the structure of the ontology alone;
4. GO is mainly an application ontology, whereas FMA is a reference ontology.

The second assessment approach was based on the notion that a pair of anatomical entities implicated in the same disease should be more related than a random pair of anatomical entities. Using HPO's annotation corpus², we were able to derive a mapping from diseases to symptoms and from symptoms to FMA concepts (see Figure 2). From this, we derived the set of pairs of *related FMA concepts*. We then extracted random pairs of other FMA concepts as the set of pairs of *unrelated FMA concepts* and performed an ROC analysis as follows:

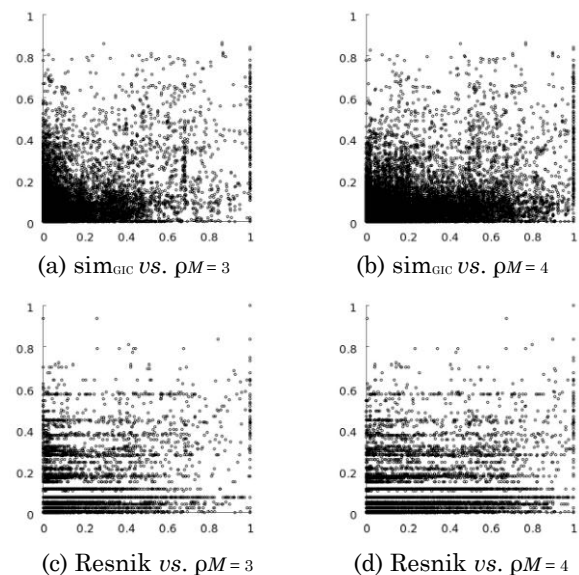


Figure 1. The correlation between sim_{GIC} and our measure of relatedness. Each point represents a mapping from GO to FMA, and its position in the graphic depends on their FMA relatedness (X-axis) and GO similarity (Y-axis) values. Correlation factors are shown in Table 1.

² We have used the MySQL dumps of HPO, available from <http://compbio.charite.de/svn/hpo/trunk/src/misc/>.

GO similarity measure	Neighborhood radius	Correlation	
		Pearson	Spearman
sim_{GIC}	$M = 3$	0.488	0.475
sim_{GIC}	$M = 4$	0.417	0.537
Resnik	$M = 3$	0.367	0.398
Resnik	$M = 4$	0.330	0.460

Table 1. The correlation coefficients corresponding to the graphics in Figure 1.

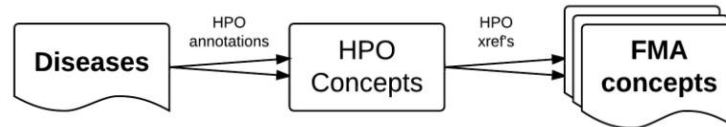


Figure 2. The work flow followed to get FMA concepts and associated diseases.

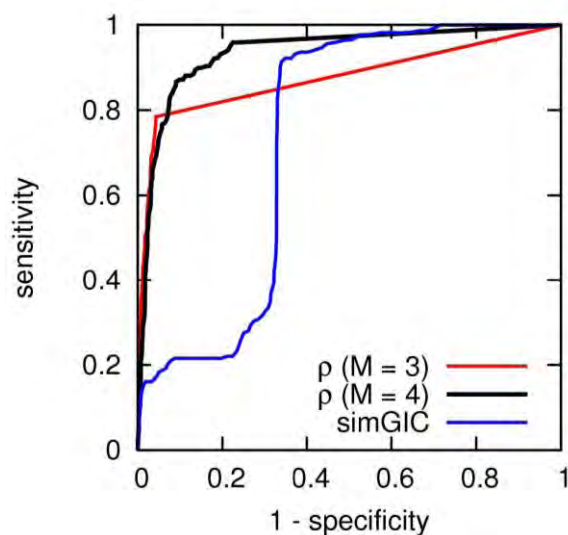


Figure 3. The ROC curves obtained from the ROC analysis. Each ROC curve is the average of 10 other ROC curves, each one produced with a different set of random unrelated pairs of FMA concepts, as described in [4] (see Algorithm 5 of that paper).

Using the relatedness measure as a score between each pair, we can arbitrarily define a threshold above which pairs should be classified as related and under which pairs should be classified as unrelated. By comparing these results with the actual related and unrelated pairs, we obtain values of sensitivity (fraction of the related concepts classified as related) and specificity (fraction of unrelated concepts classified as unrelated). Setting the threshold to values from 0 to 1, we can draw a “sensitivity vs. (1 – specificity)”, or ROC, curve [4]. These

curves are presented in Figure 3. For comparison purposes, we have also implemented the sim_{GIC} measure to FMA, according to [14].

As is evident from the figure, the best performing measure was FMA’s relatedness measure with $M = 4$, since high values of sensitivity are obtained without compromising specificity. The main difference between the measures with $M = 4$ and $M = 3$ is that the former has more resolution power in that it can differentiate between concepts 8 relations apart, whereas for the latter, concepts with a path distance greater than 6 have a relatedness value of 0.0. Additionally, for a threshold of 0.0, all pairs are classified as related (sensitivity = 1 and specificity = 0). Given the lower resolution of the $M = 3$ measure, there are a lot more pairs with relatedness value of 0.0, resulting in the straight line. With a larger value of M ($M \geq 5$), it would be possible to increase the resolution even further, at the cost of time of execution. However, this would be reflected only in the less related concept pairs, those that are more distant to one another.

Furthermore, to illustrate our assertion that semantic *similarity* is not always appropriate, consider the performance of sim_{GIC} in the same figure, which demonstrates the superiority of relatedness measures over semantic similarity, at least when applied to ontologies where a wide number of relationship types is used.

Other validation approaches are being considered, including a correlation between

relatedness and co-occurrence of concepts in a corpus, and asking experts in the area (physicians) to score pairs of anatomical concepts based on relatedness. This, however, must take into account that their background knowledge may differ significantly, e.g., a cardiologist and a physician specialized in infectious diseases may have different points of view concerning the relatedness of “Heart” and “Lung”.

5 Conclusions

With the advent of biomedical ontologies and its generalization among several fields of research, an increasing amount of ontology-based applications are emerging which leverage on the knowledge encoded in the ontologies as a way of processing and deriving new knowledge and filtering results. A measure of relatedness between ontology concepts is of the utmost importance to these applications. Here we presented a measure that is general enough that it can be applied to most extant biomedical ontologies. It is based on the concept of relevant neighborhood and relevance factors, and can accommodate the needs of particular applications by fine tuning its parameters. For example, by giving different weights to different relationship types, the measure can give more importance to some neighbors than others. Another advantage of the method is that it can incorporate external knowledge, through appropriate relevance factors, but it is not required to do so.

The concept of relevant neighborhood introduced in this work is also a bridge to other methodologies, particularly in allowing the use of ontology mappings to define wider neighborhoods that draw not only from a specific ontology but from related ontologies as well, as long as a mapping of some sort exists between the ontologies. For example, cross-references can be used for this effect.

One possible application of this measure is to improve Information Retrieval systems where resources (such as datasets, web pages and documentation) are fully- or semi-automatically annotated, both when the user is searching from keywords or trying to find resources related to a given input. Projects like the Epidemic Marketplace [9] or the RICORDO effort to integrate clinical informa-

tion [7], will consequently benefit from this measure.

Finally, a preliminary analysis was performed on FMA, and the results show that this is a valid method to measure relatedness between biomedical concepts. We expect to successfully apply the measure to other ontologies in the future, with focus on ontologies that may also be valuable to the epidemiological field.

Acknowledgments

The authors want to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807) and the FCT for the financial support of the PhD grant SFRH/BD/ 69345/2010 and the Multiannual Funding Programme.

References

1. Au, A., Li, X., Gennari, J.H.: Differences Among Cell-structure Ontologies: FMA, GO, & CCO. In: AMIA Annual Symposium Proceedings. vol. 2006, pp. 16–20. American Medical Informatics Association (2006)
2. Cao, S.L., Qin, L., He, W.Z., Zhong, Y., Zhu, Y.Y., Li, Y.X.: Semantic search among heterogeneous biological databases based on gene ontology. *Acta biochimica et biophysica Sinica* 36(5), 365–70 (2004)
3. Collier, N., Goodwin, R.M., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K., Dien, D.: An ontology-driven system for detecting global health events. *Proceedings of the 23rd International Conference on Computational Linguistics* pp. 215–222 (2010)
4. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. *Machine Learning* 31, 1–38 (2004)
5. Ferreira, J.D., Couto, F.M.: Semantic Similarity for Automatic Classification of Chemical Compounds. *PLoS Computational Biology* 6(9), e1000937 (2010)
6. Godzik, A., Jambon, M., Friedberg, I.: Computational protein function prediction: are we making progress? *Cellular and molecular life sciences* 64(19-20), 2505–11 (2007)
7. Hunter, P., Coveney, P., de Bono, B., Diaz, V., Fenner, J., Frangi, A., Harris, P., Hose, R., Kohl, P., Lawford, P., et al.: A vision and strategy for the virtual physiological human in 2010 and beyond. *Philosophical Transactions of the Royal*

- Society A: Mathematical, Physical and Engineering Sciences 368(1920), 2595 (2010)
8. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., Robinson, P.N.: Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics* 85(4), 457–64 (2009)
 9. Lopes, L., Silva, F., Couto, F., Zamite, J., Ferreira, H., Sousa, C., Silva, M.: Epidemic marketplace: an information management system for epidemiological data. *Information Technology in Bio-and Medical Informatics, ITBAM 2010* pp. 31–44 (2010)
 10. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19(10), 1275–1283 (2003)
 11. Othman, R.M., Deris, S., Illias, R.M.: A genetic similarity algorithm for searching the gene ontology terms and annotating anonymous protein sequences. *Journal of Biomedical Informatics* 41, 65–81 (2008)
 12. Patwardhan, S., Pedersen, T.: Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together* pp. 1–8 (2006)
 13. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics* 40(3), 288–99 (2007)
 14. Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E.N., Falcão, A.O., Couto, F.: Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* 9 Suppl 5, S4 (2008)
 15. Pesquita, C., Faria, D., Falcão, A.O., Lord, P.W., Couto, F.M.: Semantic similarity in biomedical ontologies. *PLoS computational biology* 5(7), e1000443 (2009)
 16. Rosse, C., Mejino Jr., J.L.V.: The foundational model of anatomy ontology, pp. 59–117. Springer (2008)
 17. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. *Genome Biology* 6(5), R46 (2005)
 18. Taylor, I.W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., Wrana, J.L.: Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology* 27(2), 199–204 (2009)
 19. Wu, X., Zhu, L., Guo, J., Zhang, D.Y., Lin, K.: Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic acids research* 34(7), 2137–50 (2006)

Aligning the Parasite Experiment Ontology and the Ontology for Biomedical Investigations Using AgreementMaker

Valerie Cross¹, Cosmin Stroe², Xueheng Hu¹, Pramit Silwal¹, Maryam Panahiazar³,
Isabel F. Cruz², Priti Parikh³, Amit Sheth³

¹Computer Science and Software Engineering Department Miami University, Oxford, OH, USA

²ADVIS Laboratory, Department of Computer Science University of Illinois at Chicago, IL, USA[†]

³Kno.e.sis Center, Department of Computer Science and Engineering Wright State University Dayton, OH, USA

Abstract. Tremendous amounts of data exist in life sciences along with many bio-ontologies. Though these databases contain important information about gene, proteins, functions, etc., this information is not well utilized due to the heterogeneous formats of these databases. Therefore, ontology alignment (OA) is now very critical for life science domains. Our work utilizes AgreementMaker for OA and describes results, difficulties faced in the process, and lessons learned. We aligned two real-world ontologies, the Parasite Experiment Ontology (PEO) and the Ontology for Biomedical Investigations (OBI). The former is more application-oriented and the latter is a reference ontology for any biomedical or clinical investigations. Our study led to several enhancements to AgreementMaker: annotation profiling, mapping provenance information, and tailored lexicon building. These enhancements, which are applicable to any OA system, greatly improved the alignment of these real world ontologies, producing 90% precision with 60% recall from the BSM^{lex+}, the Base Similarity Matcher, and 57% precision with 67% recall from the PSM^{lex+}, the Parametric String Matcher, both using lexicon lookup for synonyms. The mappings obtained through this study are posted on BioPortal for public use.

Keywords: ontology alignment, biomedical ontologies, ontology profiling, mapping provenance, lexicons.

1 Introduction

Ontology alignment (OA) is a well-recognized need for bioinformatics [10] and biomedical researchers. Currently around 260 bio-ontologies exist on the NCBO BioPortal¹ and a

number of databases exist that contain information about genes and their sequences and functions, proteins and pathway information. This knowledge, all related but modeled with heterogeneous ontologies, if better connected would greatly benefit researchers. OA addresses this challenge by identifying semantically identical or related entities in different ontologies. The resulting alignments can then be used for exchanging data and information.

Over the past decade sophisticated algorithms to improve OA have been developed. The Ontology Alignment Evaluation Initiative² (OAEI) [8] is a coordinated international effort providing standard methods for assessing the performance of OA systems. It has facilitated the advancement of OA techniques with its

[†] The ADVIS Laboratory is partially supported by NSF Awards IIS-0513553 and IIS-0812258 and by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL) contract number FA8650-10-C-7061. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

¹ <http://bioportal.bioontology.org>

² <http://oaei.ontologymatching.org>

standard set of cases and evaluation methods that developers can use to learn from and improve their OA systems. The 2010 OAEI challenge consists of 4 different tracks including one subtrack to align two anatomy ontologies, the Adult Mouse Anatomy (MA) and the NCI Thesaurus Human Anatomy (HA). This subtrack is most relevant to our research.

This paper describes the process of, lessons learned from, and results of aligning two real-world biomedical ontologies, the Parasite Experiment Ontology³ (PEO) and the Ontology for Biomedical Investigations⁴ (OBI). These two present an interesting but different scenario from the OAEI anatomy subtrack. The PEO is being collaboratively developed as part of an NIH funded project to develop and deploy an ontology-driven semantic problem-solving environment for parasite research [7]. The PEO models experiment details and provenance information of experimental data. It is an application-oriented and more specific domain ontology than OBI, which describes biomedical investigations. In contrast, the OAEI Anatomy Track ontologies have many common entities since both describe the same domain (i.e., anatomy) for different species (mouse vs. human). Mapping more specific ontologies, like PEO, to a more general ontology, like OBI, is important to helping establish a common point of reference, which can serve to foster cooperation and interoperability among researchers. Due to different scope of PEO and OBI, it may not be possible to have PEO absorbed into OBI therefore being able to create mappings between the two ontologies becomes essential. Given the explosion of metadata available on the web, this study to align two related ontologies in the biomedical field has the potential to impact not just the biomedical field but also any research field using semantic web technologies.

Aligning these ontologies requires selecting a suitable OA tool. Examining the 2010 OAEI results showed that AgreementMaker [2-4] ranked first of the nine OA systems used on the anatomy test case and performed successfully in other tracks. Since AgreementMaker had performed very well on the anatomy subtrack, had good developer support, and was readily available for our use and/or modification, it was

selected. Aligning PEO and OBI exposed the need for more flexible and configurable OA algorithms. To address this need, the features of annotation profiling, mapping provenance information, and tailored lexicon building were developed in the ADVIS Laboratory, added to AgreementMaker and experimentally validated through the PEO and OBI alignment process.

In what follows, the two ontologies, an overview of AgreementMaker, the alignment process, and the results of this study are described.

2 Ontologies

The OBI, a part of the Open Biological and Biomedical Ontologies (OBO) Foundry,⁵ describes biological and clinical investigations (e.g., designs, protocols, instrumentation). It supports the integration of experimental data across various domains such as transcriptomics, proteomics, and metabolomics through its broader scope and controlled vocabulary. The OBI incorporates concepts from the Information Artifact Ontology⁶ (IAO) and also uses annotation properties defined in the IAO.

The PEO is currently not a part of the OBO Foundry, but is found in the NCBO BioPortal. It models provenance information of experiment protocols used in parasite research and other experiment details to support annotation and querying of parasite experiment data and other databases. It references the Parasite Lifecycle Ontology⁷ (PLO). Both PEO and OBI are represented in OWL but differ greatly in size and structure (110 vs. 3060 classes) while the OAEI MA and HA are more similar (2744 vs. 3304 classes).

3 AgreementMaker

The AgreementMaker OA system has many useful features including a well designed user interface, a diverse selection of matching algorithms (matchers), and mapping quality metrics to filter and combine the results of its matchers into a final best alignment. AgreementMaker provides an extensible architecture permitting new matching or

³ <http://bioportal.bioontology.org/ontologies/42093>

⁴ <http://www.obofoundry.org/cgi-bin/detail.cgi?id=obi>

⁵ <http://www.obofoundry.org>

⁶ <http://code.google.com/p/information-artifact-ontology/>

⁷ <http://bioportal.bioontology.org/ontologies/39544>

weighting algorithms to be easily integrated and adjusted based on their performance. The user can easily evaluate, compare, and combine different strategies and matching results using its interface.

The matchers fall into two main categories: concept-based, which employ multiple string similarity measures, and structural, which search for shared patterns in the hierarchical structure of the ontologies. Our work used the Base Similarity Matcher (BSM), the Advanced Similarity Matcher (ASM), the Parametric String-based Matcher (PSM), and the Vector-based Multi-Word Matcher (VMM). The BSM calculates the similarity between two concepts by comparing all the strings associated with those two concepts including the concept name, label, and comments. PSM is also a string-based matcher but more complicated since it uses a substring measure and an edit distance measure. VMM compiles a virtual document for every concept of an ontology by concatenating the strings of related concepts and annotations, transforms the resulting strings into TFIDF vectors, and computes the similarity between those vectors using the cosine similarity measure [2]. AgreementMaker version 0.22 extended these string-based matchers by integrating two lexicons:

- (1) the Ontology Lexicon, built from synonym and definition annotations existing in the ontologies themselves, and
- (2) the WordNet Lexicon, created by starting with the ontology lexicon and adding any non-duplicated synonyms/definitions found in WordNet.

The Ontology Lexicon for each ontology (source and target) is built starting from a list of all the concepts (classes and properties) defined in an ontology. We then iterate through the list and inspect the definition of the concepts for synonym and definition annotations (*hasSynonym* and *hasDefinition* respectively). If these are found, they are added to the Ontology Lexicon entry for that concept.

The WordNet Lexicon is built starting with a previously built Ontology Lexicon. We then perform a WordNet lookup for every entry in the Ontology Lexicon and add any non-duplicated synonyms and definitions to that concept's entry in the WordNet Lexicon. It must be noted that the Ontology Lexicon and

the WordNet Lexicon are kept in separate data structures. While the Ontology Lexicon contains information directly defined in the ontology, the entries in the WordNet Lexicon can contain ambiguous information. This ambiguity must be taken into account by the matching algorithms.

The matchers using the lexicons in their algorithms are annotated with a *lex* superscript, as in BSM^{lex} , PSM^{lex} , and VMM^{lex} [4]. The Linear Weighted Combination (LWC) matcher [3] produces a single combined alignment by using mapping quality measures [3, 5] to choose the best mappings from each matcher.

4 Aligning PEO with OBI

The alignment process between the PEO and OBI is performed on the import closure of the resources, i.e., taking into account all the files each imports. The first alignment used AgreementMaker (version 0.22) with the 2010 OAEI anatomy configuration and produced only two mappings due to an inconsistency in entity descriptions of the PEO and OBI. PEO's URIs use a textual fragment identifier (e.g., <http://knoesis.wright.edu/ParasiteExperiment.owl#transfection>), while OBI's entities use numerical identifiers (e.g., http://purl.obo.library.org/obo/OBI_0600060). The PEO's use of the `rdfs:label` field does not follow the specification guidelines since when PEO happens to use this field (only 19.1% of the PEO classes have a label), it contains a PLO identifier. For the OBI, on the other hand, it uses the `rdfs:label` field to contain a descriptive string on almost 100% of its classes. The comment field for the PEO is used on 99% of its classes and typically provides a definition. The OBI only uses the comment field on about 4% of its classes. Although some common annotations exist between them, either PEO or OBI has low coverage. For example, OBI has high coverage for label annotations while PEO has high coverage for comment annotations. This heterogeneity and matchers aligning the same annotations to each other (i.e., class ID with class ID, label with label, etc.) resulted in almost no alignment.

The OAEI ontologies, in comparison, both use label and *hasRelatedSynonym* annotations and have descriptive local names. Experimental

results show lexicon synonyms and definitions greatly benefited the MA and HA alignment [4]. Their usage for the PEO to OBI alignment could also be beneficial.

As PEO and OBI do not rely on the same metadata set to describe their entities, *annotation profiling* was implemented to allow users to select and combine different annotations of the source or target ontology for the alignment. Several synonym-like annotations exist but are not found across the ontologies. The issue becomes how to match synonym annotations between the two ontologies and how to handle the mismatch in the usage of identifier, label, and definition fields. Figure 1 illustrates this feature with PEO as the source

and OBI as the target. All existing annotations in the ontologies are shown. The user selects which to use for aligning. For our experiments, the BSM and ASM matchers take advantage of annotation profiling.

Evaluating mapping results requires laboriously searching the ontologies to find descriptive names, labels, definitions, etc. This task motivated implementing the mapping provenance feature. Mapping provenance information can be automatically generated on a mapping-by-mapping basis for matchers supporting this feature. The provenance information can be interactively viewed, saved to the alignment result, and later imported.

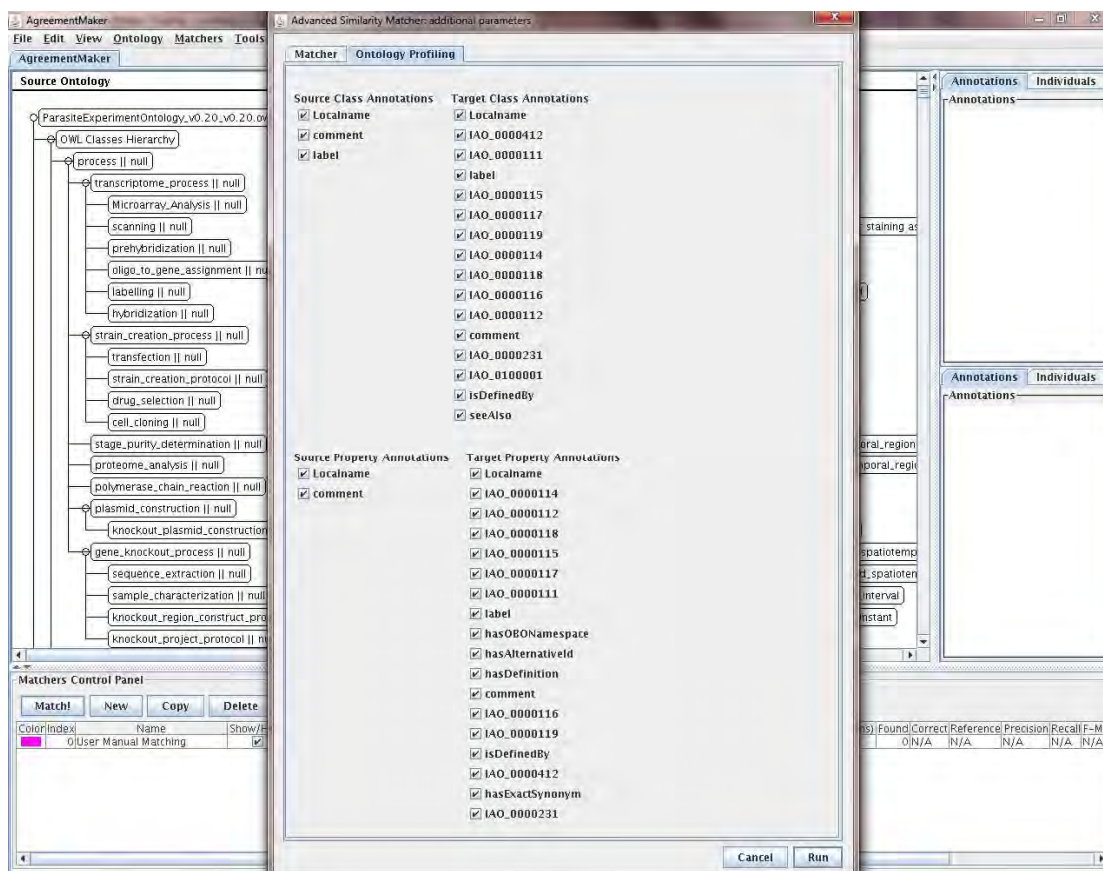


Figure 1. Profiling interface added to AgreementMaker.

```
<Cell> <entity1 rdf:resource="http://paige.ctegd.uga.edu/ParasiteLifecycle.owl#organism"/>
<entity2 rdf:resource="http://purl.obolibrary.org/obo/OBI_0302722"/>
<measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0</measure>
<relation></relation>
<provenance>sim("organism", "organ") = 1.0</provenance>
</Cell>
```

Figure 2. An example of mapping provenance information.

The provenance information, an example shown in Figure 2, provides descriptive information for both entities that may come from the names, labels, definitions, and their other associated annotations. Given provenance information, a user does not have to look up what entity OBI_0302722 actually represents but can easily see it is “organ.” Provenance capabilities are provided for ASM, BSM, PSM and VMM.

Finally, since AgreementMaker’s previous lexicon builder used a fixed name for the synonym and definition annotations (hasSynonym and hasDefinition), it was modified to exploit the synonym annotations in PEO and OBI by having the user choose the annotation names used to create the lexicons. For synonyms, OBI does not use hasSynonym but uses IAO annotation properties IAO_0000111 (“editor preferred term”) and IAO_0000118 (“alternative term”); they serve the same function as synonyms for the OBI. The PEO does not use synonyms but uses the comment annotation for a definition in most cases.

5 Experimental Results, Evaluation, and Discussion

Our domain expert evaluated the possible mappings between the entities of the PEO and OBI ontologies and produced a set of mappings. The expert provided a confidence score in the range (0.0 1.0] for each mapping. A mapping with a confidence score of 0.8 or higher is considered a correct mapping and included in the reference alignment. A total of 30 PEO to OBI reference mappings were produced. The number of mappings is low since the PEO is a more specific ontology than OBI, and only their overlapping concepts can be mapped. This reference alignment and the data for these experiments can be found on the Kno.e.sis website (http://wiki.knoesis.org/index.php/Parasite_Experiment_ontology) in the section “Alignment of PEO and OBI.” Each matcher was evaluated against the reference alignment to compute precision, recall, and F-measure. Given a reference alignment R and a computed alignment A , the precision of alignment A is calculated as

$$\text{Precision}(A, R) = \frac{|A \cap R|}{|A|}$$

and the recall is calculated as

$$\text{Recall}(A, R) = \frac{|A \cap R|}{|R|}$$

where $|A|$ represents the number of mappings in alignment A . The F-Measure is the harmonic mean of precision and recall and is calculated as

$$\text{F-Measure}(A, R) = \frac{2 \cdot \text{Precision}(A, R) \cdot \text{Recall}(A, R)}{\text{Precision}(A, R) + \text{Recall}(A, R)}.$$

Once the annotation profiling feature was implemented for the ASM, an alignment using all the annotations declared in each ontology was produced. The reasoning was that considering all the available annotation information in the matching process would lead to the best possible result, labeled ASM_{ALL} in Table 1. It shows an overall inconclusive alignment, due to low precision and medium recall. This experiment shows that matching ontology entities without discriminating between their annotations is not a viable approach – unless their annotations are semantically compatible, as seen for the OAEI ontologies.

After experimenting with ASM_{ALL}, we decided to use only the most useful and compatible annotations, reasoning this approach should give better results. Since the ASM computes alignments using a string matching similarity more suitable for short strings and compound words, the next experiment labeled ASM_{SYN} used only the synonym annotation properties declared in the ontologies. For example, the local names of the OBI ontology were not used since they are mostly ID numbers. Instead for the OBI, IAO_0000111, IAO_0000118, and label were used for class annotations and *hasExactSynonym* was added to these for the property annotations. The resulting alignment contained fewer mappings but only reduced the number of correct mappings by one mapping, leading to a 13% increase in precision while losing only 3% recall. Our reasoning was correct; however, using only a string matching algorithm was not enough to match the ontologies.

Next, the lexicons based on the modification to lexicon building process as previously

described were incorporated. All synonym and definition annotations in the ontologies were selected for use in the building of the lexicons. The BSM^{lex+} , PSM^{lex+} , and the VMM^{lex+} use the user customized lexicons in the matching process.

As suspected, the lexicons greatly improved the alignment quality. BSM^{lex+} achieves high precision with good recall; a similar performance was observed when matching the OAEI ontologies. PSM^{lex+} further improves recall by applying more sophisticated string matching algorithms. However more incorrect mappings are produced. VMM^{lex+} , which uses definition annotations, found two mappings but only one was correct. All other matchers found the same correct mapping.

With these promising individual matcher results, the next experiment combined these individual alignments into one final alignment. LWC combined the $ASMSYN$, PSM^{lex+} , VMM^{lex+} , and BSM^{lex+} into one alignment result using the “local confidence” quality metric [3] and a mapping selection threshold of 0.5.

The LWC alignment in Table 1 has the best recall but cannot avoid including incorrect mappings, leading to lower precision. The LWC combines alignments by applying a quality weight to each mapping in the input alignment; if correct mappings are only slightly better than other incorrect mappings, the combined alignment can be less precise. Although the input matchers produce good alignments, improvement is needed to better discern correct from incorrect mappings. This result, not observed for the OAEI ontologies, is most likely due to the high level of

heterogeneity between PEO and OBI annotations. Aligning the PEO and OBI, as real-world test cases, showed our lexicon based matching algorithms greatly improve alignment results, but more domain specific lexicons are needed to aid in the disambiguation of very similar entities.

Matcher	Precision	Recall	FMeasure
ASM_{ALL}	0.25	0.53	0.34
ASM_{SYN}	0.38	0.50	0.43
BSM^{lex+}	0.90	0.60	0.72
PSM^{lex+}	0.57	0.67	0.62
VMM^{lex+}	0.50	0.03	0.06
LWC	0.49	0.70	0.58
Combined	0.26	0.80	0.39

Table 1. Precision, Recall and FMeasure results for each alignment experiment.

Finally, to gauge LWC performance in producing a better alignment than just a simple combination of the input alignments, the input mappings to LWC were examined for overlap. Figure 5 shows that LWC usually can choose most of the correct mappings from each matcher. To further evaluate LWC, we manually produced a “Combined” alignment consisting of all distinct mappings produced by the six experiments, labelled “Combined” in Table 1. The Combined alignment was then compared with the reference alignment. The Combined alignment is only 10% better than the one produced by LWC in terms of recall but precision is 23% lower. This result shows that LWC can indeed discriminate between correct and incorrect mappings because of its use of an intrinsic quality measure [3], albeit not perfectly. Research may be needed to develop a more robust quality measure.

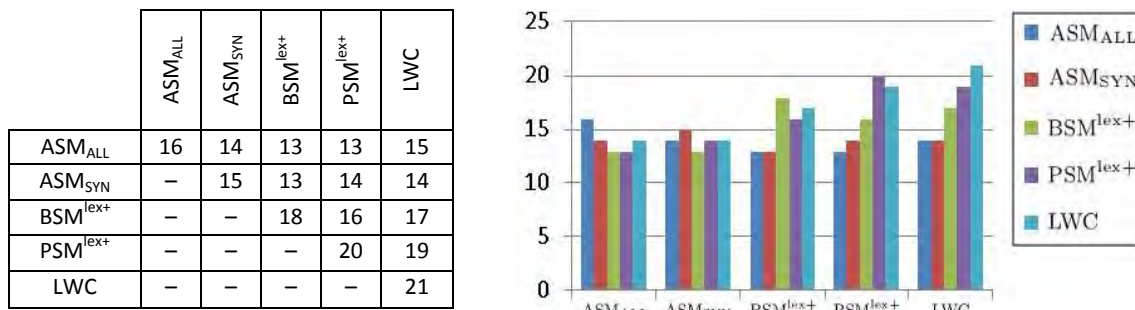


Figure 3. The number of overlapping correct mappings for each pair of matchers shown in tabular and graphical formats, showing that LWC combination of multiple matching algorithms produces a better alignment than a standalone matching algorithm.

6 Conclusions and Future Research

Aligning the PEO and OBI exposes the problem of heterogeneous annotations in ontologies. This problem can be managed by increasing the flexibility of the state of the art matching algorithms. Our implementation of annotation profiling, mapping provenance information, and custom lexicons contribute greatly to providing this flexibility.

AgreementMaker's past approach of extending matching algorithms using lexicons [4] has also been validated since the best results are produced by matchers using lexicons (e.g., BSM^{lex+} in Table 1). However, including more lexicons such as UMLS [1] needs to be investigated in order to achieve even better results. More lexicons would allow matchers to better disambiguate entities and, thus, improve the combination of the matching results.

Our current approach to managing heterogeneity relies on user selection of relevant annotations for the matching process. Annotation profiling and mapping provenance information support a domain expert in this process; however, research is needed to automatically identify semantically compatible annotations, possibly by applying established ontology evaluation metrics [6]. The heterogeneity present in real-world ontologies must be further addressed so that the OA process can foster cooperation and interoperability between researchers and organizations.

Acknowledgements

We thank Dr. Jie Zheng, postdoc at University of Pennsylvania and one of the primary developers of OBI for her initial support and help with mappings.

References

1. O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database-Issue):267–270, 2004.
2. I. F. Cruz, F. Palandri Antonelli, and C. Stroe. AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. *PVLDB*, 2(2):1586–1589, 2009.
3. I. F. Cruz, F. Palandri Antonelli, and C. Stroe. Efficient Selection of Mappings and Automatic Quality-driven Combination of Matching Methods. In Shvaiko et al. [9].
4. I. F. Cruz, C. Stroe, M. Caci, F. Caimi, M. Palmonari, F. Palandri Antonelli, and U. C. Keles. Using AgreementMaker to align ontologies for OAEI 2010. In Shvaiko et al. [8].
5. C. Joslyn, P. Paulson, and A. M. White. Measuring the Structural Preservation of Semantic Hierarchy Alignment. In Shvaiko et al. [9].
6. M. Mochol and A. Jentzsch. Towards a Rule-Based Matcher Selection. In A. Gangemi and J. Euzenat, editors, *EKA*, volume 5268 of *Lecture Notes in Computer Science*, pages 109–119. Springer, 2008.
7. S. S. Sahoo, D. B. Weatherly, R. Mutharaju, P. Anantharam, A. P. Sheth, and R. L. Tarleton. Ontology-Driven Provenance Management in eScience: An Application in Parasite Research. In R. Meersman, T. S. Dillon, and P. Herrero, editors, *OTM Conferences (2)*, volume 5871 of *Lecture Notes in Computer Science*, pages 992–1009. Springer, 2009.
8. P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, M. Mao, and I. Cruz, editors. *Proceedings of the 5th International Workshop on Ontology Matching (OM2010)* collocated with the 9th International Semantic Web Conference (ISWC-2010) Shanghai, China, November 7, 2010, volume 689 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.
9. P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, N. F. Noy, and A. Rosenthal, editors. *Proceedings of the 4th International Workshop on Ontology Matching (OM-2009)* collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009, volume 551 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.
10. H. Tan, V. Jakoniene, P. Lambrix, J. Aberg, and N. Shahmehri. Alignment of Biomedical Ontologies Using Life Science Literature. In E. G. Bremer, J. Hakenberg, E.-H. Han, D. P. Berrar, and W. Dubitzky, editors, *KDLL*, volume 3886 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2006.

A Case Study of ICD-11 Anatomy Value Set Extraction from SNOMED CT

Guoqian Jiang¹, Harold R. Solbrig¹, Robert J.G. Chalmers²,
Kent Spackman³, Alan L. Rector², Christopher G. Chute¹

¹Mayo Clinic College of Medicine, Rochester, MN. U.S.A.

²University of Manchester, Manchester, U.K.

³International Health Terminology Standards Development Organisation, Copenhagen, Denmark

Abstract. The 11th revision of the International Classification of Diseases (ICD-11) intends to derive its “ontological component” from external ontologies (e.g. SNOMED CT). One of the core value sets is an ICD-11 anatomy chapter. The objective of the present study is to develop and evaluate approaches to value set extraction from SNOMED CT for the ICD-11 anatomy use case. We investigated a number of resources comprising SNOMED CT base terms, the anatomical term mappings between ICD-O (ICD for Oncology) and SNOMED CT, the CORE Problem List Subset of SNOMED CT, and the SNOMED CT stated form in Web Ontology Language (OWL). We used the Manchester OWL module extraction tool and its extension in Protégé 4.1. We proposed and evaluated four semi-automatic value set extraction strategies based on different clinical contexts and discussed their implications in terms of domain coverage, granularity and clinical usefulness from both technical and clinical perspectives.

Keywords: ICD-11, SNOMED CT, Web Ontology Language (OWL), Ontology modularity, Value set Definition

1 Introduction

The 11th revision of the International Classification of Diseases (ICD) was officially launched by the World Health Organization (WHO) in March 2007 [1]. The WHO has sought to reuse existing ontologies such as SNOMED CT for value set definition. One of the core value sets being developed is for anatomical site, defined by WHO as “the most specific level of the topographical location or the anatomical structure where the health-related problem can be found relevant to the condition” [2].

In this context, a value set is a uniquely identifiable set of valid values that can be associated with a defined set of ICD entities. Typically, value sets can be drawn from pre-existing coding schemes such as SNOMED CT by constraining the value selection based on a logical expression (e.g. all sub-codes of the code “breast cancer”). Generating clinically meaningful value sets in a (semi-) automatic way from a terminology/ontology service has been challenging for the community, in part

due to the lack of 1) formal linkage to clinical context patterns that act as constraints in defining a concept domain; 2) techniques for automatically linking values to their appropriate concept domains, and 3) tools based on formal language such as the Web Ontology Language (OWL) [3]. To deal with some of these challenges, a number of research and standardization efforts are being undertaken, including HL7 Common Terminology Services II specification [4], Mayo’s LexEVS 6.0 implementation on a value set definition service [5], and Manchester’s new OWL API 3 [6] which contains a set of modularization tools [7].

In the present work, we performed a case study of ICD-11 anatomy value set extraction from SNOMED CT. We propose four semi-automatic value set extraction strategies based on different clinical context patterns. We evaluated the strategies and discuss their implications in terms of domain coverage, granularity and clinical usefulness from both technical and clinical perspectives.

2 Background

2.1 ICD-11 and its Content Model

WHO has embraced a broadened set of use cases to drive ICD-11 development [8]. The purpose of the ICD-11 content model is to present the knowledge that underlies the definitions of an ICD entity [2]. Table 1 illustrates that “Body System/Structure Description” is one of 13 main parameters for describing an ICD category.

1	ICD Entity Title
2	Classification Properties
3	Textual Definitions
4	Terms
5	Body System/Structure Description
6	Temporal Properties
7	Severity of Subtype Properties
8	Manifestation Properties
9	Causal Properties
10	Functioning Properties
11	Specific Condition Properties
12	Treatment Properties
13	Diagnostic Criteria

Table 1. The ICD-11 content model main parameters

2.2 SNOMED CT and its Concept Model

SNOMED CT is the most comprehensive clinically oriented medical terminology system. It is owned and maintained by the International Health Terminology Standard Development Organization (IHTSDO) [9]. The IHTSDO and the WHO signed a collaborative agreement in July 2010, which essentially establishes SNOMED CT as the core of the ontological component of ICD [10].

For its representation of anatomy, SNOMED CT has adopted a variant of the “Structure-Entire-Parts (SEP)” triple mechanism developed by Hahn and Schulz [11-12]. Fig. 1 shows a representation example for “skin structure of face”, in which “skin structure of face”, “entire skin of face” and “skin of part of face” forms a SEP triple.

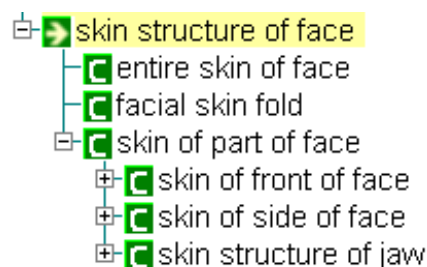


Figure 1. An example of SEP triple representation for “73897004 skin structure of face”.

SNOMED CT has a Clinical Finding concept model, in which a set of attributes has been specified to define Clinical Finding concepts [13]. Two of these attributes, “Finding Site” and “Associated Morphology,” are allowed values taken from the hierarchy under “123037004 Body structure”. Table 2 shows the two anatomy-related defining attributes.

Defining Attribute	Subsumed Attribute	Allowable Values
[FINDING SITE]		[Anatomical of acquired body] 442083009 (<<)
[ASSOCIATED MORPHOLOGY]		[Morphologically abnormal structure] 49755003 (<<)

Table 2. Two anatomy related attributes specified in SNOMED CT Clinical Finding concept model

3 Materials and Methods

3.1 Materials

For this study we used:

- 1) The topographical term mappings from SNOMED CT to ICD-O Topography provided by the 20100731 International Release of SNOMED CT;
- 2) A subset of SNOMED CT anatomy “base terms” extracted by IHTSDO from its complete set of anatomical concepts [14];
- 3) A subset of “Clinical Finding” concepts extracted from SNOMED CT stated form after conversion into Web Ontology Language (OWL) using the Perl script provided by the original distribution;
- 4) The entries of the type “finding” or “disorder” extracted from a download of the UMLS CORE Problem List Subset of SNOMED CT [15]

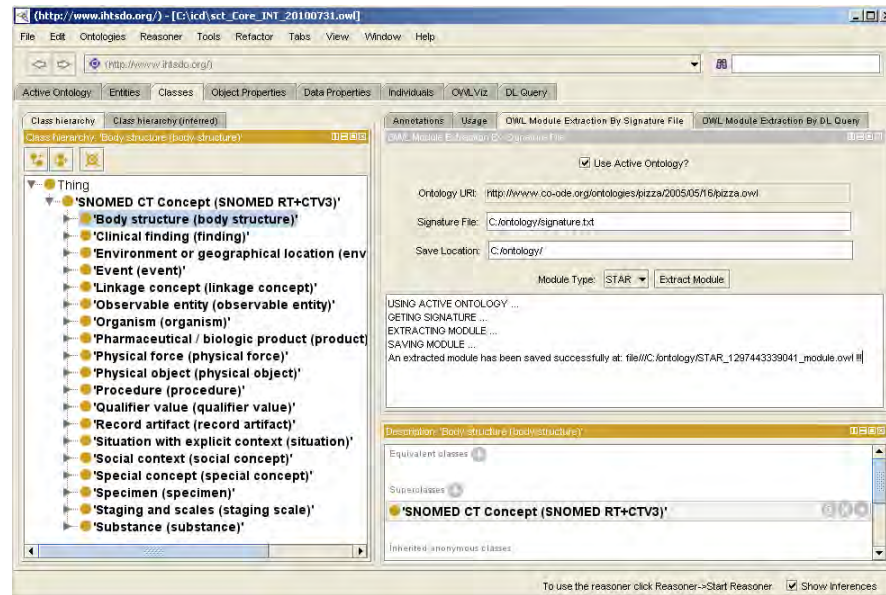


Figure 2. A screenshot of the SNOMED CT module extraction tool as a Protégé 4.1 plugin.

3.2 Methods

We took advantage of a Manchester SNOMED module extraction API [16, 17] and developed an extension as a Protégé 4.1 plugin [18]. In this way, we were able to load the SNOMED CT stated form as OWL into the Protégé 4.1 platform for module extraction. Fig. 2 shows a screenshot of the module extraction plugin in Protégé 4.1.

We defined four clinical context patterns for the ICD anatomy value set extraction, one for each of the sources described above. We created a signature file for each and identified concept IDs for 23,221 SNOMED CT concepts mapped to ICD-O topography, 14,871 concepts extracted from the anatomy base terms subset of SNOMED CT (note that the concept IDs are not available for a portion of base terms), 97,138 concept IDs extracted from the branch “Clinical Finding” of SNOMED CT and 5,304 concepts from the CORE Problem List subset of SNOMED CT.

With the four signature files, we generated four modules in OWL syntax using the module generation tool. For the first and second patterns, the generated modules are anatomy-specific modules, whereas for the third and fourth patterns, the generated modules are not, because the original signatures are all from the “Clinical Finding” domain. As the module extraction tool extracts all axioms relevant to

the signatures, the corresponding anatomical structures are also extracted. Once the modules were generated in the first round, we created a signature file for each module using the concept IDs extracted from the “body structure” sub-tree of each module. Using the signature files, we generated the anatomy-specific modules for the third and fourth patterns.

We evaluated the anatomy-specific modules extracted from the four patterns in three aspects: 1) domain coverage, 2) module granularity, and 3) clinical usefulness of the module. For domain coverage, we used the 287 ICD-O topographical categories as anchors to classify the concept IDs in each module. The domain coverage is measured by the ratio of the number of categories containing mappings over total number of categories (i.e. the 287 categories). For module granularity, we defined two measures: general granularity and adjusted granularity. The general granularity is measured simply by the ratio of the number of concept IDs in the module over the number of concept IDs in a control module (i.e. the module of the first pattern). The adjusted granularity is measured by the average ratio of the number of concept IDs in each of 287 categories in a module over that of the control module.

In addition, we performed a preliminary evaluation of clinical usefulness of the modules. We chose two categories out of the 287 ICD-O

categories “C44.3 skin of face” and “C44.5 Skin of trunk” in the dermatology domain. One of the authors (RC, a dermatology physician) reviewed the SNOMED CT mapping concepts to the two categories and marked those that should be considered as part of ICD-11 anatomy concepts. We measured the clinical usefulness by the ratio of the number of concepts checked (by RC’s ratings) over the number of total concepts in the two categories.

4 Results

In total, there are 31,107 concept IDs under the “123037004 Body structure” branch of SNOMED CT. We successfully extracted four anatomy-specific modules from SNOMED CT based on four different clinical context patterns. Table 3 shows the number of concept IDs in each module and their distribution. The majority of concept IDs in each module are those under “91723000 Anatomical structure”, whereas in the modules *ClinicalFinding* and *ProblemList*, a significant number of concept

IDs (1,919 and 755 respectively) are under “118956008 Morphologically altered structure”.

Fig. 3 shows the evaluation results of domain coverage, general granularity and adjusted granularity of the four extracted modules. The results indicate that compared with the *Control* module, the *BaseTerm* module and *ClinicalFinding* module reduced granularity approximately by one third and two third respectively but still keep good domain coverage, whereas the *ProblemList* module reduced granularity by about nine tenths while it lost domain coverage by about one fourth.

Table 4 and Table 5 show the evaluation results of clinical usefulness (which is based on RC’s ratings) for two target ICD-O categories. The results indicate that the *ClinicalFinding* and *ProblemList* modules had better outcome in terms of clinical usefulness, revealing that the clinical context patterns underlying the modules are effective and match more closely with the clinician’s expectations on ICD-11 anatomy.

	Total	ANS	ACS*	MAS*	AOP	Qualifier
Control	24142	23617	63	73	14	4
BaseTerm	15167	15004	55	70	24	4
ClinicalFinding	7120	5218	46	1919	2	4
ProblemList	2955	2201	23	755	0	4

ANS – Number of Concept IDs under “91723000 Anatomical structure”

ACS – Number of Concept IDs under “280115004 Acquired body structure”

MAS – Number of Concept IDs under “118956008 Morphologically altered structure”

AOP – Number of Concept IDs under “91832008 Anatomical organizational pattern”

Table 3. Number of concept IDs in each module and their distribution

*Note that some of concept IDs under ACS and MAS overlap.

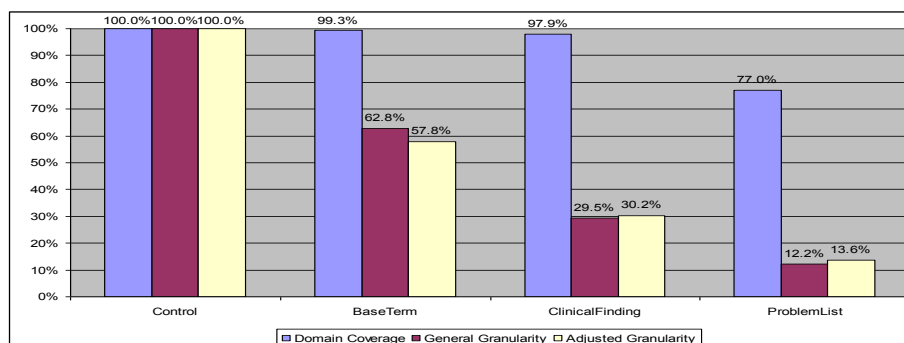


Figure 3. Evaluation results of domain coverage and granularity for four modules

	Number of Concepts Checked	Total Number	Ratio	Relative Ratio (vs. Control)
Control	17	70	24.3%	1.00
BaseTerm	17	48	35.4%	1.46
ClinicalFinding	14	34	41.2%	1.70
ProblemList	4	11	36.4%	1.50

Table 4. Evaluation results of clinical usefulness for category “C44.3 Skin of face”.
The number of concepts checked is based on RC’s ratings.

	Number of Concepts Checked	Total Number	Ratio	Relative Ratio (vs. Control)
Control	11	149	7.4%	1.00
BaseTerm	11	85	12.9%	1.75
ClinicalFinding	11	56	19.6%	2.66
ProblemList	6	20	30.0%	4.06

Table 5. Evaluation results of clinical usefulness for category “C44.5 Skin of trunk”.
The number of concepts checked is based on RC’s ratings.

5 Discussion

For the development of the ICD-11 anatomy chapter, the list of ICD-O topographical codes was a potential candidate [19] as it has been standardized in the oncology domain by WHO to describe the “Neoplasm” chapter of ICD-10. However, the list of ICD-O codes is sparse and not sufficiently detailed for many purposes.

As the IHTSDO provides mappings from SNOMED CT to ICD-O Topography as a part of the standard distribution, the mappings were naturally considered as a way to identify possible candidates for the purposes of ICD-11 anatomy. However, the problem with this idea is that the mapping is directional, and the direction is from SNOMED CT to ICD-O codes, meaning that virtually all SNOMED CT anatomy codes that could be relevant in cancer (i.e. that can be the site of a malignancy) are mapped. Moreover, the map doesn’t identify a single best SNOMED CT code for each unique ICD-O topographical code.

The entire set of SNOMED CT codes included in the map does not make a significant reduction in the size of the anatomy terminology, and leaves us with the atypical and highly unfamiliar naming of things according to the S-E-P model, such as “X structure (body structure)” and “Entire X (body structure)”. Based on the assumption that end users of the anatomy codes will want familiar names and a full set, IHTSDO created a subset of SNOMED CT which

contains 18,266 base terms, from which we extracted 14,871 concept IDs for module extraction (note that some of the former have their origin with the FMA but have no corresponding SNOMED CT concept, and thus do not have a concept ID assigned).

For the SNOMED CT Clinical Finding concept model, we consider the attributes “Finding Site” and “Associated Morphology” to be analogous to the parameter “Body structure” in the ICD-11 content model, whereas the SNOMED CT concepts under the Clinical Finding branch are analogous to the disease categories in ICD-11, although the SNOMED CT concepts are more fine-grained. We consider that the asserted anatomical structures corresponding to the Clinical Finding concepts are meaningful to be a clinical context pattern for the ICD-11 anatomy value set.

We defined quantitative measures to evaluate the four modules in terms of domain coverage, module granularity and clinical usefulness. We believe that the measures are useful to help in deciding which module extraction strategy is effective and should be considered for adoption. Note that the usefulness evaluation we performed in the present study has limited generalizability because it is based on a single reviewer’s ratings. It seems clear that we could obtain more reliable results by using more experts from diverse clinical backgrounds.

Based on our results, we suggest that the

strategy used for the module *ClinicalFinding* as a good starting point for the ICD-11 anatomy use case. The module has good domain coverage while keeping a relatively small size and better outcome on clinical usefulness. Note that the SNOMED CT anatomy base terms are still useful for providing familiar names and a full set, and are complementary to our suggested strategy here.

In addition, the upper level of the SNOMED CT anatomy may create some confusion because it has three main branches: 1) Anatomical structure, i.e. the normal anatomy, 2) Acquired body structure - mostly the results of surgery plus “scar (morphologic abnormality)”, and 3) Body structure altered from its original anatomic structure (morphologic abnormality) – the results of disease or repair. Given these different types of body structure, the question really is what is needed for ICD-11 development. The clinical experts in the WHO TAG groups, editorial boards and the clinical groups working on anatomy will need to clarify their requirements in order to determine whether one or all of these branches should be used.

Finally, the Protégé based OWL module extraction tool we developed in this study [18] has been demonstrated very useful for achieving our goal. While we mainly use an external signature file for module extraction in this study, we have extended the tool and integrated it with the Protégé DL Query plugin [20], by which a signature can be defined through a semantic query which invokes a DL (description logic)-based reasoner. We consider that this provides a powerful and easy to use feature to define and extract a domain specific value set.

In conclusion, we performed a case study for ICD-11 anatomy value set extraction from SNOMED CT. We proposed four different clinical context patterns for the purpose of generating clinically meaningful value sets for ICD-11 anatomy. We evaluated the value sets in terms of domain coverage, granularity and clinical usefulness by defining quantitative measures, which provide effective metrics for helping us to select an approach for satisfying the ICD-11 anatomy use case.

References

1. WHO. Revision of International Classification of Diseases (ICD): <http://www.who.int/classifications/icd/ICDRevision/en/index.html>.
2. WHO. ICD-11 Alpha – Content Model Reference Guide: <http://icat.stanford.edu/>.
3. Pathak J, Jiang G, Dwarkanath SO, Buntrock JD, Chute CG. LexValueSets: an approach for context-driven value sets extraction. AMIA Annu Symp Proc. 2008 Nov 6:556-60.
4. CTS2 wiki: <http://informatics.mayo.edu/cts2>.
5. The LexEVS 6.0 URL: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexEVS_Version_6.0.
6. The OWL 3 API URL: <http://sourceforge.net/projects/owlapi/>.
7. Sattler U, Schneider T, Zakharyashev. Which kind of module should I extract? In: DL Home 22nd International Workshop on Description Logics, July 27-30, 2009, Oxford, UK.
8. Chute CG. Distributed biomedical terminology development: from experiments to open process. Yearb Med Inform. 2010:58-63.
9. The IHTSDO URL: <http://www.ihtsdo.org/snomed-ct/>. Last visited at February 15, 2011.
10. Agreement between IHTSDO and WHO: <http://www.who.int/classifications/AnnouncementLetter.pdf>.
11. Hahn U, Schulz S, Romacker M; Partonomic reasoning as taxonomic reasoning in medicine. Proc 16th National Conf Artificial Intelligence & 11th Innovative Applications of Artificial Intelligence (AAAI-99/IAAI-99); 271-276.
12. Schulz S, Hahn U, Romacker M; Modeling anatomical spatial relations with description logics. 2000; AMIA Fall Symposium, 799-783.
13. The IHTSDO. SNOMED CT Clinical Terms User Guide. July 2010 International Release.
14. SNOMED CT anatomy data file: <https://csfe.aceworkspace.net/sf/go/doc3132?nav=1>.
15. The CORE Problem List Subset of SNOMED CT Download: http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html.
16. Manchester SNOMED CT Module Extraction Tool: <http://owl.cs.manchester.ac.uk/snomed/>.
17. Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT Hierarchies in practical applications. JAMIA. 2011. (in press).
18. The Protege4.1. Plugin for SNOMED CT Module Extraction URL: <https://sites.google.com/site/ontologymodularit/>.
19. The International Classification of Diseases for Oncology (ICD-O) URL: <http://www.who.int/classifications/icd/adaptations/oncology/en/>.
20. The Protégé DL Query Plugin. <http://protégé.wiki.stanford.edu/wiki/DLQueryTab>.

The HL7 Approach to Semantic Interoperability

Jobst Landgrebe¹, Barry Smith²

¹Cognotekt LLC, Germany; ²National Center for Ontological Research, University at Buffalo, NY, USA

Abstract. Health Level 7 (HL7) is an international standards development organisation in the domain of healthcare information technology. Initially the mission of HL7 was to enable data exchange via the creation of syntactic standards for point-to-point messaging. For some 10 years, however, HL7 has increasingly conceived its mission as one of creating standards for semantic interoperability in healthcare IT on the basis of its ‘version 3’ (v3) family of standards. Unfortunately, v3 has been marked since its inception by quality and consistency issues, and it has not been able to keep pace with recent developments either in semantics and ontology or in computer science and engineering. To address these problems, HL7 has developed what it calls the ‘Services-Aware Interoperability Framework’ (SAIF), which is intended to provide a foundation for work on all aspects of standardization in HL7 henceforth and which includes HL7’s Reference Information Model as general purpose upper ontology. We here evaluate the SAIF in terms of design principles that must be satisfied by a semantic interoperability framework, principles relating both to ontology (static semantics) and to computational behaviour. We conclude that the SAIF fails to satisfy these principles.

1 Introduction

Health Level 7 (HL7) has been producing standards with the goal of enabling interoperability in the health care domain since 1987. Since the end of the 1990s, HL7 has been developing its ‘version 3’ (v3) set of standards, which have been associated with the goal of achieving *semantic* interoperability, defined as the ability of two or more computer systems to communicate information and have that information properly interpreted by a receiving system in the same sense as was intended by the transmitting system (adapted from [14]).

The central pillar of HL7’s attempts to achieve this goal was the HL7 Reference Information Model or ‘RIM’, initially conceived by HL7 as a modeling language specific to the healthcare domain. HL7’s idea was that all data models and message types would henceforth be derived from the RIM in a way that would serve to counteract the formation of local dialects affecting v2 implementations. Unfortunately this strategy has encountered a number of difficulties. The RIM is highly complex and its foundations are idiosyncratic both ontologically [31] and technically [28]. Technically, the RIM is based on a blend of UML [23] and XML 1.0 [4], that is not formally specified, and it deviates in important ways from accepted norms of

modeling. This has made the RIM difficult to use, both for humans and for computer systems.

By 2007, the HL7 leadership had come to realise that v3 was not achieving significant uptake. At the same time, the point-to-point messaging-based integration approach traditionally followed by HL7 was itself proving too narrow to meet the demands of the IT industry, where interoperability paradigms rooted in the theory of distributed computing, in ontologies, and in net-centric data standards, were being used with considerable success.

To address these issues, the HL7 leadership initiated in 2007 its Services-Aware Interoperability Framework (SAIF) [12], which is now implemented at the National Cancer Institute as part of the recently scrutinised caBIG initiative [7–9], and which HL7 sees as providing a framework to achieve ‘working interoperability’ in the E-Health Domain [12].

Our approach in what follows is to identify principles which have been applied in other areas of IT in achieving semantic interoperability, and to evaluate the SAIF in light of these principles. Our analysis concludes that the SAIF fails to satisfy these principles with regard to both architecture and computational behaviour as well as ontology and information modeling.

2 Methods

2.1 Principles for Semantic Interoperability Frameworks

We define a semantic interoperability framework (SIF) as a set of guidelines describing how to create a system architecture for distributed computing [1], whose implementation would allow multiple independent systems to exchange, correctly interpret, reuse and aggregate data provided only that these systems conform to the architecture defined in the framework.

By analyzing the workings of two well established engineering frameworks in which these benefits are realized – IEEE 1471 [15] and the Reference Model of Open Distributed Processing (RM-ODP) [17] – we were able to identify certain basic principles that, we believe, must be satisfied if a proposed framework is to be capable of such realization. We then validated this list of principles by examining further examples, listed in Table 1, of standards specifying guidelines for achieving distributed computational behaviour. We know of no relevant frameworks or guidelines which would not satisfy the principles here listed.

Our principles fall into two groups. First are certain high-level principles which must be satisfied in the authoring of frameworks for any kind of computer architecture. Second are principles specific to the requirements of those architectures that are designed to support semantic interoperability.

General Framework Principles

1. Domain boundaries – a SIF must be associated with a single, coherent domain in which it is declared to be valid, and it must define explicitly the boundaries of this domain.
2. Knowledge reuse – a SIF must build upon established scientific and technical knowledge and best practices in order to minimize the barriers to adoption and, by drawing on what has been validated in use, thereby maximize the likelihood of success.

3. Level of abstraction – a SIF must be sufficiently abstract that it can support systems implemented across a broad spectrum of technical alternatives.

Interoperability Framework Principles

1. Enterprise requirements – a SIF must describe how to identify and formalise the requirements associated with those specific business processes, functions, workflows, and desired outcomes which systems specified and engineered on its basis are intended to support [15, 17].
2. Information model and ontology – a SIF must separate information content from the means by which this content is exchanged, and it must specify on two levels how information content will be structured in such a way as to ensure that it can be communicated successfully from one system to another, first on the level of data through an information model and second on the level of linking data to reality through an ontology [33, 17].
3. Computational model – a SIF must allow the specification both of the behaviour of the single systems conforming to the framework and of the interactions between these systems. [17]
4. Architecture framework and conformance model – a SIF must describe how a system architecture must be structured if it is to be conformant to the framework; thus it must include a list of necessary system components and describe how these in their turn need to be structured in order to conform to the framework [15].

Satisfaction of these principles is not, by any means, sufficient for achieving semantic interoperability: the principles are guidelines only. We believe, however, that they represent necessary conditions. They identify, in effect, component features which belong to the very definition of semantic interoperability. Thus if the principles are violated, semantic interoperability is unachievable.

3 Results and Discussion

3.1 Description and Evaluation of the SAIF

The SAIF attempts simultaneously to address three ‘interoperability paradigms’, namely *messaging* – which is the traditional method for exchanging data in healthcare environments; *documents* – a term here referring to HL7’s XML-based Clinical Document Architecture standard [10]; and *services*, where interoperability relates to the ability of two or more computer systems to communicate information and request specific behavior over service interfaces in such a way that meaning is preserved while details of the internal service implementation are hidden from the user.

To realize semantic interoperability it is necessary to address both structure (of information transmitted) and behavior (of interoperating systems). We shall address the former – which is where ontology comes into play – in the section about the Information Framework below. For the moment we note only that HL7 addresses structure in the same way in its treatment of all three paradigms, namely in terms of the RIM.

When it comes to behavior, however, matters are not so simple. Indeed, the document paradigm does not address behavior at all: a Document, for HL7, is simply a packaged set of information with certain metadata attached. HL7 has tried to address behavior in its messaging infrastructure with its ‘dynamic model’, which can be compared to modeling the way information is exchanged in a walkie-talkie conversation, and which many in the HL7 community now accept as being outmoded [11]. Only in its treatment of the services paradigm do we enter a domain with a level of technical maturity that can in principle allow the kind of specification of the behavior of modern computer systems that is needed to realise interoperability.

Components of the SAIF. As specified by HL7 [12], the SAIF consists of four components: the Behavioural, Information and Governance Frameworks as well as the Enterprise Conformance and Compliance Framework. We will deal with each of these in turn.

The Information Framework (IF). This consists of (i.) a set of five principles and (ii.) a set of

artefacts [12].

i. The first three IF principles essentially state that data, information, knowledge and their respective representations are different from each other and that this difference should be respected in a SAIF-conformant framework. These principles are certainly true, but they are also trivial and it is not clear how they relate to the remainder of the SAIF documentation. The fourth IF principle calls for a separation of what HL7 designates as ‘formal concept representation’ (the realm of clinical terminologies) from what it calls ‘clinical linguistics’ (the realm of natural language). It is not clarified what this distinction really means and why it is made.

The fifth IF principle asks for traceability from ‘information concepts to organizational/technical concepts and patterns.’ Traceability is essential for any engineering framework. One must, for example, be able to prove that certain user requirements justify a certain technical and financial effort. The standard approach to requirements traceability is a formalisation of the requirements which must be satisfied to meet the needs of an organisation followed by a systematic derivation of a stack of engineering artefacts from these requirements. In the SAIF, however, traceability is defined in a direction exactly opposite to that of all the approaches with which we are familiar – by proceeding from information models to organisation models. It is not clear to us how on this basis traceability can be realized.

Taken together, the five IF principles seem to us to have the character of a preamble. They do not provide context for the subsequent sections of the SAIF specification nor are they referenced by it.

ii. *Artefacts of the IF* include: domain analysis models, reference information models, domain information models, serializable information models, localized information models, types – classes, attributes, data types, semantic types, and vocabulary (including value sets and value set bindings to attributes).

It is clear from a number of passages (for example where the subject of discussion is the HL7 ‘cascade’ by which information models are ‘scoped’ and ‘specialised by constraint’) that when mention is made here and elsewhere in the SAIF documentation to ‘reference information models’ then it is the HL7 RIM that

is intended. Most significant in this respect is the final sentence of section 2.4 of [12]:

In fact, if the reference information model is abstracted to a coarse level of entities and the relationships of those entities through roles to the actions that they somehow participate in then it can be conceptually applicable to any information domain or sector. One can think of a reference information model as an upper ontology that describes the static semantics of all possible real world information (emphasis added).

Something new, here, is that the RIM (in effect) is now being described, and advocated, by HL7 as an ‘upper ontology’. This is significant because the RIM constrains everything in reality to fall under only three types, namely: Act (meaning roughly: an intentional action performed, for example by a clinician), Role (as, for instance, the nurse role), and Entity (meaning roughly: something containing molecules as parts, including persons). This attenuated repertoire of types causes problems when the RIM is called upon to serve as an ontology framework ‘that describes the static semantics of all possible real world information.’ Indeed HL7 v3 has from the very beginning faced problems even when applied to health-related information – for example when it comes to dealing with diseases. For the latter are neither Acts nor Entities nor Roles [31].

Similarly, the RIM leaves no room to represent key items in medical reality of other types, such as drug interactions, infections, accidental falls, processes of inflammation – all of which are identified by HL7 as Acts of Observation.

These, and other, technical problems with the RIM make v3 models hard to author and confusing to read, and they bring well-known problems in achieving interoperation with standard vocabularies such as SNOMED CT [29]. Above all, the large number of message-interchange-related (and thus neither ontology- nor information-model-related) attributes built into the core of the RIM imply a failure to separate content from the means for exchanging content [27]. This leads to problems both in data serialisation and in the creation of cleanly computable structures, and explains in turn the lack of practical adoption of the RIM.

The RIM suffers, too, from the idiosyncratic way in which XML has been used to represent both it and its derived models. This is because the RIM exposes design properties of

the XML technology to the end user, thereby violating the SIF design principle of abstraction. As an example, consider the way in which metadata structures are related to class attributes in the RIM classes in a nested fashion in direct correspondence with the underlying nested XML structures. This makes it hard for a modeler to understand the usage and meaning of the class attributes and prevents fluent and intuitive modeling with HL7 v3.

In addition to insisting on the RIM both as modeling language and (now) as upper ontology, the IF’s description of its artefacts shows also that HL7 plans to adhere to its existing approach as concerns overall modeling and vocabulary usage. Rather than seizing the opportunity to use well established information modeling practices that are in use across the information technology industry in order to simplify the tasks faced by healthcare information modelers, IF’s treatment of artefacts serves rather to justify the existing, complex and in many ways idiosyncratic HL7 v3 approach. (As an illustration of this phenomenon, consider this statement from [12] on Localized Information Models:

A localized model may be derived from a larger serializeable model. For instance, a serializeable model may be localized with constraints on datatypes or more refined concept domains than its parent. In this case, the localized model is a logical model and may be used as a serializeable model. It may also be derived by refinement and constraint of a portion of a domain analysis model and may not be serializeable, in which case it is still a conceptual model like its parent, or may be constrained in such a way from the domain information model that it can be serialized in which case it would be a logical model.)

The Behavioral Framework (BF). As we have seen, achieving interoperability on the service paradigm requires specification of the behaviour of computational units. But this requirement is not at all specific to healthcare. Rather, it is an overarching problem in computer science [2], and has given rise to a number of established frameworks, of which the prime examples, as already mentioned, are IEEE 1471 [15] and RM-ODP [17] (further examples are provided in Table 1). Given that, as we shall see, SAIF in any case uses many elements of RM-ODP, one has to question HL7’s rationale for defining its own health-care-specific hybrid rather than simply using

the BF defined by RM-ODP.

The BF documentation [12] falls into two parts. First, sections 3.1 to 3.7 provide a framework for the specification of behaviour drawing on terms and definitions from RM-ODP and the ODP Enterprise Language [16]. The remaining sections of the document present ideas on BF implementation and on what are called ‘correspondences’, a term not clearly defined, but which seems to refer to links between the BF and HL7’s now obsolete ‘dynamic model’ [11]. Terms like ‘compound binding’s contract correspondence’ seem to have no precedent in the literature on computational behaviour and in the pertinent standards. Comparison of successive versions of the BF documentation makes it clear that a shift has been occurring in HL7 towards incorporating RM-ODP more directly, and that this has provided some incremental benefits. It is not clear what the corresponding sections add beyond RM-ODP, since no references are provided to any previous work or to efforts at empirical validation in existing implementations.

The Governance Framework (GF) has the aim of describing roles and responsibilities of the persons involved in implementing architectures conforming to interoperability specifications. We do not analyse the GF here as the analysis would go beyond the scope of this communication.

Finally, there is the SAIF’s *Enterprise Conformance and Compliance Framework (ECCF)*, whose most conspicuous feature is that it contains as part the SAIF architecture specification. System architecture, in other words, is addressed under the heading of conformance. In all existing interoperability frameworks known to us, it is architecture which subsumes conformance. To see what results from SAIF’s approach, consider its treatment of traceability to requirements definitions, which SAIF sees as one component of conformance. A conformance and compliance framework as normally conceived defines a procedure whereby it is possible to follow a set of requirements from specification down to realisation in bits and bytes and thereby test for fulfillment. Conformance hereby presupposes architecture. Because neither the SAIF nor any other HL7v3 standard describes how requirements should be specified, HL7 is unable to specify how the content against which a trace

is to be established should be constructed, and so there is no possibility of requirements traceability assessment.

Further problems derive from the core element of ECCF, the so-called Specification Stack, a matrix shown in Figure 1, which is a combination of RM-ODP and of OMG’s model driven architecture (MDA) stack [21]. The ECCF claims that specifying artefacts (essentially specifications, code and test case documentation) in terms of how they instantiate the ECCF matrix provides a recipe for creating system architectures enabling semantic interoperability. Unfortunately, however, the ECCF matrix has a construction problem. Mathematically speaking, it is singular, which means that its column vectors are not linearly independent. This is because the viewpoints identified by RM-ODP (designated as ‘dimensions’ by the SAIF) already cover the full space of system engineering alternatives [17]. It therefore does not make sense to combine RM-ODP with MDA as if this would yield a matrix providing further coherent alternatives. Certainly the RM-ODP viewpoints can be combined with different weights to create different levels of abstraction. But this does not mean that RM-ODP and MDA can be put together to yield a cross-tabulation. The attempt to do so results in a set of instructions that cannot be followed and implemented. Since, if we leave the mentioned problems aside, the useful portions of the ECCF prove once again to be a re-packaging of definitions from the RM-ODP, one must question once more whether there is value that is being added.

3.2 Summary Evaluation of SAIF

We can now summarize our results by examining the degree to which the SAIF realises the framework principles outlined above.

General Framework Principles

1. *Domain boundaries.* By defining in its preamble its scope as “Working Inter operability ... between business objects, components, capabilities, applications, systems and enterprises” [12] thus with no restriction to the health domain the SAIF indicates that it has arrogated to itself a domain that essentially covers the whole of IT.

2. *Knowledge reuse.* SAIF's definition of its own scope violates, too, the principle of knowledge reuse, raising the question of why HL7 should have included areas within its remit – such as the specification of computational behaviour – which have been already successfully covered by other efforts.
3. *Level of abstraction.* The RIM provides no clean separation of the logical modeling layer from the underlying implementation technology. The RIM in its current form is thus not technology independent – that is, it is not a logical artefact which can be expressed using multiple technologies – because it is permeated by HL7-specific XML design principles.

Level of Abstraction	Domain	Examples of Established Standards
Enterprise Level	Interactions and collaborations within and between enterprises and systems	OASIS Reference Architecture Foundation for Service Oriented Architecture [6], openGroup SOA Reference Architecture [34] (for overview see [19])
System Level	Orchestration (service interactions guided by a controlling master unit)	Business Process Execution Language, Business Process Modelling Notation [18, 35], Unified Modelling Language 2 activity diagrams [25]
	Choreography (service interactions not moderated by a controller but effected by computational units which are peers of each other)	Process algebra (one flavour of which is better known as π -calculus) [2, 22, 13], Occam [20], a process algebra-based language used in parallel computing.
Individual Service Level	Service interface, service behaviour	SOA-ML [24], Z notation [32]

Table 1. Examples of standards specifying guidelines for achieving distributed computational behaviour

ECCF	Enterprise Dimension "Why" - Policy	Information Dimension "What" - Content	Computational Dimension "How" - Behavior	Engineering Dimension "Where" - Implementation	Technical Dimension "Where" - Deployments
Conceptual Perspective	<ul style="list-style-type: none"> ✓ Inventory of <ul style="list-style-type: none"> o Use Cases, Contracts o Capabilities-Services o Stakeholders o Non-Functional Requirements o Methodologies/Processes o Policies & Regulations o Business Objectives ✓ Business Mission, Vision, Scope 	<ul style="list-style-type: none"> ✓ Inventory of <ul style="list-style-type: none"> o Domain Entities o Stakeholders, Roles o Activities o Associations o Information Requirements o Information Models <ul style="list-style-type: none"> • Conceptual • Domain 	<ul style="list-style-type: none"> ✓ Inventories of <ul style="list-style-type: none"> o Capabilities-Components o Functions-Services ✓ Requirements <ul style="list-style-type: none"> o Accountability, Roles o Functional Requirements, Profiles, Behaviors, Interactions o Interfaces, Contracts ✓ Functional Service Specifications 	<ul style="list-style-type: none"> ✓ Inventory of <ul style="list-style-type: none"> o SW Platforms, Layers o SW Environments o SW Components o SW Services o Technical Requirements o Enterprise Service Bus ✓ Key Performance Parameters 	<ul style="list-style-type: none"> ✓ Inventory of <ul style="list-style-type: none"> o HW Platforms o HW Environments o Network Devices o Communication Devices ✓ Technical Requirements
Logical Perspective	<ul style="list-style-type: none"> ✓ Business Policies ✓ Use Case Specifications ✓ Governance ✓ Implementation Guides ✓ Technology Neutral Standards ✓ Wireframes of <ul style="list-style-type: none"> o Architectural Layers o Components and Associations ✓ Contracts 	<ul style="list-style-type: none"> ✓ State Variables ✓ Information Models <ul style="list-style-type: none"> o Localized o Constrained o Project ✓ Vocabularies ✓ Value Sets ✓ Content Specifications <ul style="list-style-type: none"> o Messages o Documents o Services 	<ul style="list-style-type: none"> ✓ State Machines ✓ Specifications <ul style="list-style-type: none"> o Use Cases, Interactions o Components, Interfaces ✓ Collaboration Participations ✓ Collaboration Types & Roles ✓ Function Types ✓ Interface Types ✓ Collaboration Scripts ✓ Service Contracts 	<ul style="list-style-type: none"> ✓ Models, Capabilities, Features and Versions for <ul style="list-style-type: none"> o SW Environments o SW Capabilities o SW Libraries o SW Services o SW Transports 	<ul style="list-style-type: none"> ✓ Models, Capabilities, Features and Versions for <ul style="list-style-type: none"> o HW Platforms o HW Environments o Network Devices o Communication Devices
Implementable Perspective	<ul style="list-style-type: none"> ✓ Business Nodes ✓ Business Rules ✓ Business Procedures ✓ Business Workflows ✓ Technology Specific Standards 	<ul style="list-style-type: none"> ✓ Schemas for <ul style="list-style-type: none"> o Databases o Messages o Documents o Services o Transformations 	<ul style="list-style-type: none"> ✓ Automation Units ✓ Technical Interfaces ✓ Technical Operations ✓ Orchestration Scripts 	<ul style="list-style-type: none"> ✓ SW Specifications for <ul style="list-style-type: none"> o Applications o GUIs o Components ✓ SW Deployment Topologies 	<ul style="list-style-type: none"> ✓ HW Deployment Specifications ✓ HW Execution Context ✓ HW Application Bindings ✓ HW Deployment Topology ✓ HW Platform Bindings

Figure 1. ECCF specification stack [12]

Interoperability Framework Principles

1. *Enterprise requirements.* The principle to the effect that a SIF must describe how to determine and formalise requirements associated with pertinent business processes is not addressed by HL7, which has no process or formalism to define requirements [26]. This means that SAIF-conformant information models may be created without traceable link to formalised requirements, which would preclude the achievement of interoperability.
2. *Information model and ontology* is violated because, for the reasons indicated in the IF section above, the RIM is neither a useable information model nor a functional ontology.
3. *Computational model* asserts that a SIF must allow the specification of both the behaviour of and of the interactions between conformant systems. This principle is addressed by the SAIF's Behavioral Framework. As described above, the BF is neither comprehensive nor consistent. Overall, it would have been better to simply assemble a consistent selection of the results of relevant mainstream efforts and to indicate how each may be applied within a healthcare specific context, following a methodology already proven, for example, in the defense [3] and logistics [30] industries.
4. *Architecture framework and conformance model*, too, is not adequately addressed. As we saw, the normal relationship of architecture and conformance is inverted; such architectural guidelines as are provided are in consequence superficial. In addition, the core of the ECCF, the specification stack, fails to fulfill its role for the reasons given above. Overall, the ECCF reflects an interpretation and usage of RM-ODP that is not in accordance with the latter's ISO specification.

4 Conclusion and Recommendations

It has to be acknowledged that, with the SAIF initiative, HL7 is attempting to evolve its

standards in the direction of contemporary interoperability paradigms. However, in light of the above analysis, we believe that the SAIF still has serious defects when measured in these terms. For SAIF to make a positive difference, we recommend that it be replaced by an approach that is based on a more fundamental reassessment of HL7 v3 that is in compliance with the interoperability design principles presented above. This does not require the development of new frameworks and methodologies: almost all of the needed components, including requirements formalisms (e.g. [5]), can be taken from existing standards and frameworks. Those components which genuinely need to be healthcare specific fall under the heading of what we have called 'information model and ontology'. To obtain useful results here, we recommend that HL7 adopt a tested upper-level ontology framework and an efficient, scientifically well founded, modular and composable domain-specific modeling language. We also recommend that HL7 replace its current approach to decision-making where it involves the use of balloted standards when addressing technical issues; when facing the sorts of complex challenges encountered in the realization of semantic interoperability, this is not an appropriate mechanism to constrain how engineers do their work. We recommend further that, like OMG and some other standards development organizations, HL7 adopts the requirement that working implementations should be provided before a standard can be published and that these implementations should be subject to a process of technical validation. Given its impressive expertise in the clinical domain we are confident that HL7 can produce real value; but to this end it should focus on what it can do best: specifying requirements describing those healthcare processes that need to be supported by Information Technology in the healthcare domain.

References

1. Attiya, H., Welch, J.: Distributed Computing: Fundamentals, Simulations, and Advanced Topics. Addison Wesley (2004)
2. Baeten, J.C.M., Basten, T., Reniers, M.A.: Process Algebra: Equational Theories of Communicating Processes, Cambridge Tracts in

- Theoretical Computer Science, vol. 50. Cambridge University Press, Cambridge, UK (2010)
3. Bardo, B. (ed.): Proceedings of Fifth Annual Conference of the Systems Engineering for Autonomous Systems Defence Technology Centre. SEAS DTC (2010)
 4. Bray, T., Paoli, J., Sperberg-McQueen, C. (eds.): Extensible Markup Language (XML) 1.0. W3C, <http://www.w3.org/TR/1998/REC-xml-19980210> (1998)
 5. Cockburn, A.: Writing Effective Use Cases. Addison Wesley (2003)
 6. Estefan, J., Laskey, K., McCabe, F., Thornton, D. (eds.): Reference Architecture Foundation for Service Oriented Architecture. OASIS, <http://docs.oasisopen.org/soa-rm/soa-ra/v1.0/soa-ra-cd-02.pdf>, 1st edn. (2009)
 7. Goldberg, P.: Contracting disputes and software bugs plague NCI's 200m bioinformatics venture. The Cancer Letter 37(8) (2011)
 8. Goldberg, P.: How NCI's plans for software giveaway sank in scientific and legal disputes. The Cancer Letter 37(9) (2011)
 9. Goldberg, P.: Scrutiny of caBIG Exposes Conflicts, Bonanza for Contractors. The Cancer Letter 37(11) (2011)
 10. HL7 (ed.): Clinical Document Architecture R2. HL7, <http://www.hl7.org/implement/standards/cda.cfm> (2005)
 11. HL7 (ed.): HL7 Version 3 Standard: Core Principles and Properties of Version 3 Models, Release 1. HL7, www.hl7.org/ctl.cfm?action=ballots.home (2011)
 12. HL7 Architecture Board (ed.): HL7 Service Aware Interoperability Framework: Canonical Version, Release 1 (Unique Ballot ID: SAIF CANON R1 I1 2011MAY). HL7, www.hl7.org/ctl.cfm?action=ballots.home (2011)
 13. Hoare, C.: A model for communicating sequential processes. Prentice Hall (1985)
 14. IEEE (ed.): IEEE Standard Computer Dictionary. A Compilation of IEEE Standard Computer Glossaries. IEEE (1991)
 15. IEEE (ed.): IEEE Std 1471-2000, Recommended Practice for Architectural Description of Software-intensive Systems. ISO, <http://www.isoarchitecture.org/ieee-1471/> (2000)
 16. ISO/IEC (ed.): Information Technology – Open Distributed Processing – Reference Model – Enterprise Language. ITU-T X.911 ISO/IEC 15414 (1998)
 17. ISO/IEC (ed.): Information technology Open Distributed Processing. ISO 10746, 1st edn. (1998)
 18. Jordan, D., Evdemon, J. (eds.): Web Services Business Process Execution Language Version 2.0. OASIS, <http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpelv2.0-OS.html> (2007)
 19. Kreger, H., Estefan, J. (eds.): Navigating the SOA Open Standards Landscape Around Architecture. Object Management Group, OpenGroup, OASIS, <http://www.opengroup.org/onlinepubs/7699909399/toc.pdf> (2009)
 20. McEwan, A., Schneider, S., Ifill, W., Welch, P. (eds.): Communicating Process Architectures 2007 -WoTUG-30. IOS Press (2007)
 21. Miller, J., Mukerji, J. (eds.): MDA Guide Version 1.0.1. OMG, 1.0.1 edn. (2003)
 22. Milner, R.: Communication and Concurrency. Prentice Hall (1989)
 23. Object Management Group (ed.): Unified Modeling Language. OMG, 1.0 edn. (1997)
 24. Object Management Group (ed.): Service oriented architecture Modeling Language (SoaML). OMG, 1.0 edn. (2009)
 25. Object Management Group (ed.): Unified Modeling Language, Superstructure. OMG, 2.2 edn. (2009)
 26. Oemig, F., Blobel, B.: A communication standards ontology using basic formal ontologies. Stud Health Technol Inform. 156, 105–113 (2010)
 27. openEHR (ed.): openEHR-Clinical mailing list archives. openEHR, <http://www.openehr.org/mailarchives/openehr-clinical/msg00210.html> (2005)
 28. openEHR (ed.): openEHR and HL7v3. openEHR, <http://www.openehr.org/206OE.html> (2007)
 29. Rector, A., Qamar, R., Marley, T.: Binding ontologies and coding systems to electronic health records and messages. Applied Ontology 4, 51–69 (2009)
 30. Smirnov, A.V., Shilov, N.: Business network modelling: SOA-based approach and dynamic logistics case study. IJISMD 1(4), 77–91 (2010)
 31. Smith, B., Ceusters, W.: HL7 RIM: An incoherent standard. In: Medical Informatics Europe 2006. pp. 133–138 (2006).
 32. Spivey, J.: The Z Notation. A Reference Manual. Prentice-Hall, 2 edn. (1992)
 33. Spyns, P., Meersman, R., Jarrar, M.: Data modelling versus ontology engineering. SIGMOD Record (ACM Special Interest Group on Management of Data) 31(4), 12–17 (2002)
 34. The OpenGroup (ed.): SOA Reference Architecture. openGroup, <http://www.opengroup.org/projects/soa-ref-arch/>, 1st edn. (2009)
 35. White, S., Miers, D.: BPMN Modeling and Reference Guide. Future Strategies Inc., Lighthouse Point, Florida (2008).

Representing the Reality Underlying Demographic Data

William R. Hogan, Swetha Garimalla, Shariq A. Tariq

Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Abstract. Demographic data about patients, research subjects, students and trainees, physicians and other healthcare providers, and so on is extremely important for nearly every biomedical application that manages information about people. The importance of demographics extends beyond biomedical informatics and touches fundamentally nearly every software application that manages information about people. However, we show that the treatment of demographic data in current information systems is ad hoc, and current standards are insufficient to support accurate capture and exchange of demographic data. We propose a solution based on realist ontology and implemented in the Demographics Application Ontology, which draws terms from reference ontologies such as the Phenotypic Quality Ontology and the Ontology of Medically Related Social Entities. Furthermore we have created a web site that demonstrates the approach.

Keywords: Realist ontology, demographics, referent tracking

1 Introduction

Demographic data at their essence are data about people. There is no consensus set of “characteristics” that comprises demographics, but typically demographics include birth date, gender, sex, marital status, race, and ethnicity. Demographics are ubiquitous in software applications in healthcare and beyond, with obvious importance, for example, to electronic health records (EHRs), to the United States Census, and to the finance and retail industries. Demographic data are used to identify people, to make statistical comparisons of population groups, and to link records from multiple databases about one person.

At the University of Arkansas for Medical Sciences, we are studying Referent Tracking (RT) [1] as applied to EHRs. Clearly, we need to include demographic data in RT Systems (RTSs). As we shall discuss, we found current approaches to (and standards for) demographics inadequate. Therefore, we analyzed the reality on the side of the person that demographic data are about. On the basis of our analysis we propose a realist approach to demographics.

2 Preliminaries

Because there is no standard set of demographics, we had to choose one to start. Here,

we discuss our rationale for the set we chose. We also set the stage for a review of current approaches to demographics by clarifying sex vs. gender.

2.1 Demographics Addressed in this Work

We started with the set of demographics required by the “meaningful use” (MU) regulation for EHRs in the United States (U.S.) [2], because we are studying RT applied to EHRs. For “eligible providers”, this set includes preferred language, gender, race, ethnicity, and date of birth. For “eligible hospitals”, it additionally includes date and preliminary cause of death (in the event of mortality in the hospital).

We then modified this set for pragmatic reasons. We *included* sex and marital status. We included the former because current approaches confuse it with gender (as we illustrate next). We included the latter because, in the U.S. at least, marriage confers on one’s spouse broad healthcare visitation and (when one is incapacitated) decision-making rights. Providers therefore require marital status to know whose instructions to follow when the patient is incapacitated. We *excluded* preferred language, race/ethnicity, and preliminary cause of death. The first requires ontological theories of preferences and human language that are beyond our scope here. The second requires a treatment that is ongoing work, and for which

we had insufficient space here to do justice. The third we assume can be handled by ongoing work in the ontology of disease and injury (e.g., Scheuermann et al.[3], Hogan [4], Cowell and Smith [5], and Goldfain et al. [6]).

2.2 Sex vs. Gender

The World Health Organization (WHO) explains the distinction between sex and gender thusly:¹

“Sex” refers to the biological and physiological characteristics that define men and women. “Gender” refers to the socially constructed roles, behaviours, activities, and attributes that a given society considers appropriate for men and women.

Sex is biological; gender is psychosocial. In practice, the correlation between them is high, but there are transgendered and transsexual people who we will need to represent correctly in EHRs for optimal patient care and research.² Furthermore, with respect to sex, we can distinguish phenotypic and chromosomal (karyotypic) sex. Although they too correlate highly, there are individuals, for example, with an XY karyotype who are phenotypically female [7].

Driven largely by administrative requirements (i.e., billing), almost all structured healthcare data at present include only gender.³ However, as we move towards a quality-driven healthcare system, quality of EHR data will become increasingly important, and thus also will, we believe, correct distinctions among gender, phenotypic sex, and karyotypic sex.

3 Limitations of Current Approaches

Here, we review usual approaches to demo-

graphics and standards for them. We divide these approaches into three major groups: “Person table”, terminology standards, and semantic web.

3.1 The “Person Table” (or Class)

The typical approach to demographic data, especially in software that uses a relational database for persistence, is to have a “Person” table with fields for birth date, marital status, etc. Formal information models do little more than more than formalize the Person table as a class, and the fields as “attributes.” Thus, the HL7 Reference Information Model (RIM) has a Person class, which “specializes” the LivingSubject class. The latter has an “attribute” called ‘administrativeGenderCode’,⁴ and the former has “attributes” called “maritalStatusCode”, “raceCode”, etc.

The limitation of this approach is that it treats gender, date of birth, and marital status in exactly the same way. That is, to the computer the only difference between gender and date of birth is the field name and datatype: else they are the same type of entity (“attribute”) related to the person in exactly the same way (“attribute of”). The true relationship between a human being and his or her gender, marital status, etc. is implicit and obscured.

3.2 Terminology Standards

The information model approach specifies that some attributes take a coded value. For example, the HL7 RIM has the codes M, F, UN as allowed values for “administrativeGenderCode”, where UN stands for “undifferentiated”. HL7 describes UN further by saying: *The gender of a person could not be uniquely defined as male or female, such as hermaphrodite.* And here we see that HL7 is confused, as ‘hermaphrodite’ refers to anatomical considerations and therefore sex, not gender. The RIM has no attribute for sex.

Terminology standards also confuse sex and gender. SNOMED CT (SNCT) places *Male* and *Female* as children of *Finding of biological sex*, but has no representation of male and female gender. It is thus either incomplete or assumes the sex codes are sufficient for gender too.

¹ <http://www.who.int/gender/whatisgender/en/index.html>

² There was a discussion on the listserv of the Association of Medical Directors of Information Systems (AMDIS) where one physician commented that he discussed with his patient how, in the future, he would have to “flip” the patient’s gender in the registration system to ensure correct payment for prostate cancer screening (male-to-female transsexual surgery leaves the prostate in place and intact). For more on the AMDIS thread, see author WRH’s discussion at <http://hl7-watch.blogspot.com/2010/12/demographics-hl7-vs-reality-part-2.html>

³ MU also requires gender only, and not sex.

⁴ Given that the primary use case for this attribute is bed assignment, we wonder why a “bed assignment instruction” code is not more apropos.

Furthermore, it also has 2547121016 *Gender determination by chromosome analysis*. Clearly, we cannot ascertain one's socially constructed roles from chromosomal analysis; karyotype is what is meant here. The UMLS also confuses sex and gender, mapping SNCT codes for sex to UMLS concept unique identifiers for gender (e.g., *Male* to C0024554 *Male gender*).

LOINC has two codes (11882-8, 11883-6) for the sex of a fetus as observed on ultrasound, but labels the measured attribute as 'gender'. It would be hard to ascertain, even by ultrasound, the socially constructed roles of a fetus. Both codes have a "related name" of 'fetal sex'.

The NCIT has *Male gender*, but not as a subtype of *Gender*, oddly. Instead, it is a subtype of *Male*, itself a subtype of "*General qualifier*". The siblings of *Male gender* include *Male phenotype* and *XX male*. The NCIT therefore asserts that karyotype, phenotype, and gender are subtypes of an informational "qualifier" (vs. biological or social) entity.

With respect to marital status, rather than simply state "married" vs. "single", terminologies contain numerous codes that combine other information such as living arrangements (married living apart), marital history (how the most recent marriage ended), stage of life (spinster), length of marriage (newlywed), number of spouses (monogamous), etc. SNCT has 35 marital status codes, including *Eloped*, *Divorced*, *Monogamous*, *Remarried*, etc. The NCIT has a smaller set of 10 codes, but still conflates status with history (e.g., *Never married*, *Annulled*).

The common practice of such combinatorics with respect to marital status frustrates interoperability – each standard proposes a different set of combinations. An effort to harmonize marital status codes among several standards development organizations (SDOs) such as HL7 and ANSI failed because one SDO went out of business. Characteristic of the combinatory approach, there was subsequent re-opening of discussion and numerous new proposals for different standards and the reconsideration of particular combinations.⁵

⁵ As discussed on a thread of the HL7 Vocabulary Workgroup listserv, reproduced here: <http://hl7-watch.blogspot.com/2010/11/demographics-hl7-vs-reality-part-1.html>

3.3 Semantic Web

The Friend of a Friend (FOAF) project is a semantic-web-based effort to standardize information about people (and more). Just as the information model approach merely reifies tables as classes and their fields as attributes (with respect to demographics), FOAF reifies tables as Web Ontology Language (OWL) classes and table fields as relations.⁶ For example, it merely replaces the Person table with a Person class, the gender field with a gender relation, and the birth date field with a birthday relation.⁷ Worse, the range of the gender relation is the "string" data type. Thus FOAF does not support interoperability of gender data at all, as any string is compliant. The semantic web approach as embodied by FOAF thus does little for our understanding of demographics or for interoperability of demographics.

4 The Realist Approach to Selected Demographics

Our hypothesis in this work was that an ontological analysis of the reality that underlies demographic information could bring coherence to demographic data. We also anticipated a reduction in the need for demographics-specific relations such as "gender", "birthday", etc.

4.1 Dates of Birth and Death

In reality, there is a birth (death) event, and the person who was born (died) is the agent in that event. This event occurs during a particular day. We thus represent John Doe's birth date of February 26, 1981 as:⁸

```
john_doe_birth instance-of Birth9
john_doe agent-of john_doe_birth at t1
john_doe_birth occurs t1
t1 instance-of Temporal instant
t2 instance-of Temporal interval
```

⁶ In the semantic web community, a relation is usually referred to as a 'property', after OWL.

⁷ The semantic web VCard 'standard' similarly has a birthday property, with a different URI.

⁸ Translation of these statements into Referent Tracking templates is straightforward.

⁹ Relations between occurrences are not time indexed. See Smith et al. [8].

t1 **during** t2
t2 **denoted-by** '1981-02-26'

We would represent his death similarly. We use representational units (RUs) from the Advancing Clinico-Genomic Trials Master Ontology (ACGT-MO) [9] for *Birth* and *Death*, where they are correctly placed as subtypes of *occurent*.

A benefit of this representation is that we can link other information about John Doe's birth (death) to its representation as required. By contrast, the usual approaches cannot accommodate such linking because the dates are just a field/attribute/property about which we can say nothing more.

4.2 Gender

We agree with the WHO that genders are socially constructed roles. We created the Ontology for Medically Related Social Entities (OMRSE)¹⁰ in part to represent gender roles properly.

We then represent John Doe's gender as:

john_doe_gender **instance-of** *Male gender*
since t2
john_doe_gender **inheres-in** john_doe since
t2

Ceusters and Smith first suggested representing gender as a separate entity, but they did not relate it to the person [10]. We arbitrarily chose John's birth date as the time his gender began to exist. However, John's parents might have learned his sex and thus chosen a male name, purchased male baby clothes, etc. before his birth. We cannot envision a need for a more precise date, but were a better date known and considered important, we could easily use it instead of birth date.

We can also represent transgendered individuals, by stating for example that the person's gender ceased to instantiate *Male gender* and began to instantiate *Female gender* at a given time:

john_doe_gender **instance-of** *Male gender*
during t2 to t3

¹⁰ OMRSE is publicly available at
<http://code.google.com/p/omrse/through an SVN repository>

john_doe_gender **instance-of** *Female gender*
since t3

4.3 Sex

Phenotypic sex, at the level of granularity of the whole organism, is a quality. The reason is that biological sex refers to more than just reproductive organs. In humans, sex also includes differences in distribution of body hair, levels of hormones, deepness of voice, and so on. Thus, sex differences are widespread throughout the body. The Phenotypic Quality Ontology (PATO) accurately represents *Phenotypic sex* as a descendant of *Organismal quality*.

We represent John Doe's sex using the PATO URI for *Male sex* as:

john_doe_sex **instance-of** *Male sex* since t4
john_doe_sex **inheres-in** john_doe since t4

Sexual differentiation occurs early during fetal development, with notable differences around four weeks. Thus for t4, we could use a date of four weeks after an estimated date of conception. Greater precision is likely not necessary but again, we could use a more precise date if available. Our approach can also accommodate ongoing sexual development by representing, for example, Tanner stages as qualities and tracking instantiation over time as with gender.¹¹

4.4 Marital Status

Because it is the legal aspects of marriage that motivate the capture of marital status in EHRs, we focus on its contractual aspects. In the U.S. at least, federal and state governments recognize marriages by conferring upon the couple certain legal obligations and rights. The key rights of concern to healthcare involve hospital visitation and decision making.

Each individual in the marriage is a party to a marriage contract.¹² We represent this role in OMRSE; it is a subtype of *Party to a legal entity*. We then represent John Doe's "married

¹¹ We note that the usual approaches cannot even begin to cope with this problem, because they do not track sex, gender, etc as single entity.

¹² Common-law marriage implies a contract between the parties that can only be terminated as with other marriages. Even then, only 9 states still recognize it and those states frequently require evidence of mutual agreement before conferring recognition.

status” as (where t5 is the wedding date):

```
john_doe_mr_role instance-of Party to a marriage contract since t5  
john_doe_mr_role inherits-in john_doe since t5
```

What about John’s unmarried brother Jack? For an affirmative statement that Jack is single, we follow Ceusters et al. [11]:

```
jack_doe lacks Party to a marriage contract  
w.r.t. bearer-of since t6
```

If t6 is Jack’s birth date, then we have successfully captured the semantics of “Single never married” included in many terminologies, but without increasing the number of RUs (or codes) in the ontology.

If either John gets a divorce or is widowed, we can update our representation:

```
john_doe_mr_role instance-of Party to a marriage contract during t5 to t7  
john_doe_mr_role inherits-in john_doe during t5 to t7
```

In states that confer on same-sex couples obligations and rights similar to those of marriage, we represent this situation with *Party to a domestic partnership agreement* from OMRSE:

```
jim_doe_dp_role instance-of Party to a domestic partnership agreement since t8  
jim_doe_dp_role inherits-in jim_doe since t8
```

This representation also allows us to say other things about the contract. For example, to handle jurisdictional issues, we can represent the relation between the party role and the contract, and between the contract and the jurisdiction that recognizes it in OMRSE. When necessary – e.g., when a marriage in one country is not recognized by another – we can capture jurisdictional information and relate it to the person’s role. None of the usual approaches offer this flexibility.

5 The Demographics Application Ontology

Our representation uses RUs from several, different realism-based reference ontologies. To facilitate implementation in RTSs and other applications that manage demographic data, we have created the Demographics Application

Ontology (DAO).¹³ The DAO does not, and will not, create new RUs for types.¹⁴ It uses the Minimum Information to Reference an External Ontology Term approach [12] to import RUs from reference ontologies.

6 Discussion

Given ubiquity of demographics in information systems used by entities in the economy worldwide and governments, the problem of accurate representation is of critical importance. We have represented the reality underlying key components of demographic data, namely dates of birth/death, gender, sex, and marital status. In doing so, we untangled much of the web of confusion surrounding sex and gender in concept-based approaches. Although Milton was the first to recognize this confusion [13], we have demonstrated its pervasiveness throughout leading concept-based artifacts including HL7, SNCT, LOINC, NCIT, and UMLS. We also reduced marital status to the essence of the rationale for capturing it in the first place. Along the way, we have demonstrated improved flexibility of representation when additional information is required.

Our representation of demographics need not complicate data entry for users of EHRs or other applications. To illustrate this fact, we implemented our approach at <http://demappon.info/Demographics.php>. It shows a typical, web-based form for entering demographics. After submitting data, the user sees the RT templates that the RTS created behind the scenes, with detailed explanations.

Our approach has the potential to simplify standardization of demographic data. We have removed the requirement for shared field/attribute/property names such as “administrativeGenderCode” (HL7) and “gender” (FOAF). Indeed, nothing in our approach requires demographics-specific relations – we used existing relations from the Relation Ontology (RO) and Ontology of Biomedical Investigations (OBI). Thus, any application that understands these relations in addition to

¹³ Available publicly at <http://code.google.com/p/demapp-ontology/>

¹⁴ The current version does include fiat classes, a.k.a., attributive collections, as OWL individuals for the purpose of capturing race information. Our work on race is ongoing.

several RUs from PATO, OMRSE, etc. (gathered into the DAO for convenience), will correctly interpret our representations.

Finally, our approach led us to appreciate the diversity of the types of entities involved with demographic data, including qualities (sex), roles (gender and marital status), events (birth and death), etc. that form fundamental distinctions at the top levels of multiple upper ontologies. The distinction between qualities and roles is important because qualities are always exemplified when present whereas roles are not. This distinction is blurred by usual approaches to demographics.

This diversity also illustrates the requirement for application ontologies such as the DAO in the realist approach. The DAO facilitates representing demographics using RUs from reference ontologies, which represent portions of reality without respect to particular use cases. Then, to facilitate a given use case, such as demographics, the application ontology can pull together the needed RUs and relationships among them.

7 Conclusion

Despite the apparent simplicity of demographic data, and thus the expectation that they ought to be easy to standardize, few if any standards for demographics data enjoy widespread adoption. We have illustrated that usual approaches to demographics and standards for them fail to account for the reality underlying them, and that representing this reality has the potential to simplify standardization and increase the flexibility and extensibility of the representation.

Acknowledgements

This work was supported by award numbers 1UL1RR029884 and 3 P20 RR016460-08S1 from the National Center for Research Resources. The content is solely the responsibility of the author and does not necessarily represent the official views of the NCRR or NIH.

References

1. Ceusters W, Smith B. Strategies for referent tracking in electronic health records. *J Biomed Inform.* 2006 Jun;39(3):362-78.
2. Medicare and Medicaid Programs; Electronic Health Record Incentive Program, 412, 413, 422, and 495 (2010).
3. Scheuermann RH, Ceusters W, Smith B, editors. Toward an ontological treatment of disease and diagnosis. *AMIA Summit on Translational Bioinformatics*; 2009.
4. Hogan WR. Towards an ontological theory of substance intolerance and hypersensitivity. *J Biomed Inform.* 2010 Feb 10.
5. Cowell LG, Smith B. Infectious Disease Ontology. In: Sintchenko V, editor. *Infectious Disease Informatics*: Springer New York; 2010. p. 373-95.
6. Goldfain A, Smith B, Cowell L. Dispositions and the Infectious Disease Ontology. In: Galton A, Mizoguchi R, editors. *FOIS*. Amsterdam: IOS Press; 2010. p. 400-13.
7. Jorgensen PB, Kjartansdottir KR, Fedder J. Care of women with XY karyotype: a clinical practice guideline. *Fertil Steril.* 2010 Jun;94(1):105-13.
8. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. *Genome Biol.* 2005;6(5):R46.
9. Brochhausen M, Spear AD, Cocos C, Weiler G, Martin L, Anguita A, et al. The ACGT Master Ontology and its applications - Towards an ontology-driven cancer research and management system. *J Biomed Inform.* 2010 May 11.
10. Ceusters W, Smith B. Referent tracking for treatment optimisation in schizophrenic patients: A case study in applying philosophical ontology to diagnostic algorithms. *Web Semantics: Science, Services and Agents on the World Wide Web.* [doi: DOI: 10.1016/j.websem.2006.05.002]. 2006;4(3):229-36.
11. Ceusters W, Elkin P, Smith B. Negative findings in electronic health records and biomedical ontologies: a realist approach. *Int J Med Inform.* 2007 Dec;76 Suppl 3:S326-33.
12. Courtot MI, et al. MIREOT: The minimum information to reference an external ontology term. *Applied Ontology.* 2011;6(1):23-33.
13. Milton SK. Top-Level Ontology: The Problem with Naturalism. In: Varzi AC, Vieu L, editors. *Formal Ontology in Information Systems*. Amsterdam: IOS Press; 2004.

Information Models and Ontologies for Representing the Electronic Health Record

Daniel Karlsson¹, Martin Berzell², Stefan Schulz³

¹Department of Biomedical Engineering, Linköping University, Sweden

²Division of Health and Society, Department of Medical and Health Sciences, Linköping University, Sweden

³Institute of Medical Informatics, Statistics and Documentation, Medical University Graz, Austria

1 Introduction

The term “information model” is often used in biomedical informatics to refer to an aggregation of information items used to represent (what a health professional knows about) the health status of the patient, as well as planned, scheduled or carried out health care activities. As an example of this use, the International Standard ISO 12967:2009 defines the term “information object” as “information held by the system about entities of the real world, including the ODP (Open Distributed Processing) system itself, is represented in an information specification in terms of information objects, their relationships and behaviour” [1]. Thus, information models seem to closely relate to ontologies, for example as defined by Spear [2]. The aim of this paper is to add to the clarification of the relationship between information models and ontologies and to propose and discuss demarcations of these two kinds of representation.

2 Background

The representations and communication of facts, beliefs, and opinions in the field of health care, and therefore the scope of the Electronic Health Record, clearly exceeds the simple instantiation of mind-independent representational units such as provided by typical biomedical ontologies. The EHR includes observations, opinions, instructions (plans), proposals, requests, etc. Ontologies catering to these needs must necessarily go beyond the realm of the mind-independent, as the reference to entities in such discourses do not necessarily imply the existence of those entities. For instance, a record entry on a “denied tonsillectomy”, initially does not refer to any tonsillectomy in clinical reality. Nevertheless,

such a proposition should be adequately represented by referring to some artifact which includes a representational unit “tonsillectomy”.

Beale and Heard propose a Clinical Investigator Record Ontology [3] distinguishing entries in the record of various types, including:

- if something has been done or if something is yet to be realized (or not), for example distinguishing a plan from a performed activity,
- if something concerns the state of the patient or if something concerns health-care activities, for example distinguishing the blood pressure of a patient from the observation thereof, and
- if something is an assessment or opinion or if something is (more or less) directly observed, for example distinguishing the observation of a blood pressure of 150/90 from a diagnosis of stage 1 hypertension.

From this Clinical Investigator Record Ontology, the *openEHR* foundation has constructed a reference (information) model and, in a larger cooperation, a framework for representing specialisations of that reference model known as archetypes [4]. Archetypes consist of constraints on how the reference model may be instantiated in addition to the constraints of the reference model. The demarcation of the reference model and its specialisations through archetypes, called the two-level modelling approach, is based on the stability of the respective models [5]. The reference model consists of a core representation which is assumed to be stable over time and across organisations while archetypes are used to build representations where such assumptions cannot be made. The *openEHR* foundation also keeps a repository of clinical and demographic archetypes [6].

```

ELEMENT[at0008] occurrences matches {0..1} matches { -- Position
  value matches {
    DV_CODED_TEXT matches {
      defining_code matches {
        [local::
          at1000, -- Standing
          at1001, -- Sitting
          at1002, -- Reclining
          at1003, -- Lying
          at1014; -- Lying with tilt to left
          at1001] -- assumed value
        }
      }
    }
  }
}

```

Figure 1. Part of an openEHR Blood Pressure archetype specification

However, not all classes in the *openEHR* reference model are clear cut. For example, the **OBSERVATION** class represents, on the one hand, the quality which is observed in the patient, and on the other hand, the performing of the observation activity, thus causing an overlap with the **ACTION** class, which represents the performing of activities. This is however consistent with another principle of *openEHR*: that each single archetype should contain everything necessary to be clinically relevant for specified use cases. In order to assess a result of an observation it is important to have knowledge of how the observation was performed.

Traditionally, the burden of formally representing clinical practice and the health status of the patient has been divided upon a number of technical solutions. Certain parts of this domain have been represented using ontology-language solutions using different kinds of Description Logic (DL), for example the one in which SNOMED CT is represented [7], while other parts of this domain have been represented using database schemas, UML diagrams or, as described above, archetypes. This division into different representational artifacts can be seen as a “divide-and-conquer” approach to dealing with the inherent complexity of medicine and clinical practice. This division is however not clear and there is no consensus on how to make this division, although attempts has been made [8, 9].

This lack of consensus can be exemplified by looking at the (now classical) blood pressure archetype and the SNOMED CT representation of the same entity. The archetype includes a

separate slot for stating the state of the patient at the time of observation (see figure 1), for example whether the patient was standing, sitting or lying down. SNOMED CT gives (almost) equal opportunities for representation with codes like:

```

163034007|Standing blood pressure
(observable entity)|, 163035008|
Sitting blood pressure (observable
entity)|, and 163033001|Lying blood
pressure (observable entity)|.

```

The selection of representation in each specific case is up to the clinical modeller and although emerging guidelines exist [8] our own experiences tells us that such guidelines are hard to apply. Also, we have yet not found evidence of any systematic useage of such guidelines.

3 Representational Requirements of the EHR

The requirements for representing the EHR can be divided into two separate cases: runtime requirements, and design-time requirements. In runtime, the amount of data stored is typically very large and the computational complexity of individual inferences performed must be kept very low for example when doing inferencing on a population scale. In design time, when for example building archetypes, the size of the models is moderate at most. Due to the uniqueness of archetypes, ontologies can generally be partitioned, for example when doing validation through satisfiability checking

only the reference model and any specialised archetypes need to be reasoned over. Thus, inferences of greater complexity may be allowed in design time.

Earlier work has shown that it is feasible to represent information model schemas such as the *openEHR* reference model [10, 11]. Analysing the ontologies representing the reference model, the concept constructors used are concept conjunction (\sqcap), existential- (\exists), value- (\forall) and number restrictions, and nominals. Value restrictions are used for attribute data type specifications in archetypes and cardinality- and occurrence constraints in archetypes are represented using existential- and number restrictions. As shown by Baader et al., subsumption testing is intractable with respect to general TBoxes [12, 13]. Number restrictions, when added to conjunction and existential restrictions, lead to EXPTIME-completeness. Archetypes used to specialize reference model classes further add qualified number restrictions [14] to the list of required concept constructors.

Some parts of the *openEHR* reference model will require special attention, specifically what is called the *openEHR* Archetype Profile which includes some classes with special semantics. For example, the reference model class **DV_QUANTITY** includes the attribute unit to represent unit of measurement. This attribute should be consistent with an Archetype Profile constraint on kind of quantity, for example that

a meter unit is consistent with a length kind of quantity.

As an example of an *openEHR* reference model class, the **OBSERVATION** class is used to represent observations of the patient's health status, see figure 2. The **OBSERVATION** class inherits from the **CARE_ENTRY** class and adds two attributes: *data*, to represent the results of the observation and *state*, to represent the state of the patient at the time of the observation. From the **CARE_ENTRY** class the inherited attribute *protocol* is used to represent the observation procedure. The **CARE_ENTRY** class in turn inherits the attribute *subject* to distinguish which patient this aggregate of information concerns.

The **data**, **state** and **protocol** parts may be further specified in archetypes. Figure 3 shows part of an observation archetype from the *openEHR* archetype repository expressed in a DL syntax. The **data** part represents the results of the observation, in this case the blood pressure values.

Additionally, opinions, plans, and proposals are significant parts of the health record. However, as noted by Rector and Brandt as well as Schulz et al., representations of plans require a more expressive description logic than the EL profile used by most large scale DL-based ontologies including SNOMED CT [15, 16].

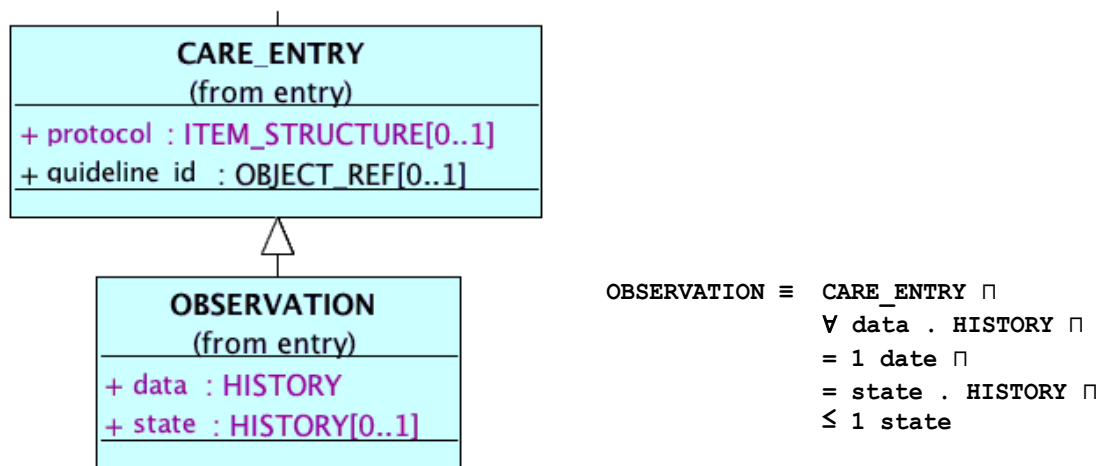


Figure 2. The *openEHR* OBSERVATION class in UML and DL


```

Blood-pressure  $\equiv$  OBSERVATION  $\sqcap$ 
     $\forall$  data . ( HISTORY  $\sqcap$ 
        ...
         $\leq 1$  items . Systolic  $\sqcap$ 
         $\leq 1$  items . Diastolic  $\sqcap$ 
        ...
    )  $\sqcap$ 
    =1 data
    ...
Systolic  $\equiv$  ( ELEMENT  $\sqcap$ 
     $\forall$  value . ( C DV QUANTITY ... )  $\sqcap$ 
    =1 value )
Diastolic  $\equiv$  ( ELEMENT  $\sqcap$ 
     $\forall$  value . ( C DV QUANTITY ... )  $\sqcap$ 
    =1 value )

```

Figure 3. Part of an *openEHR* OBSERVATION archetype in DL

Recently, the IHTSDO has presented a style guide for representing observable entities and procedures in laboratory medicine in SNOMED CT [17] influenced by, among other things, the OBO phenotypic quality ontology (PATO) [18]. Both rely on the EL profile for representation: PATO for representing qualities and the SNOMED CT observables model for representing qualities as well as how those qualities are observed and represented, indicating that both qualities and observation procedures may be represented using tractable DL.

4 Discussion

When choosing means of representation for a specific use case, there is always the trade-off between expressivity and computational complexity to consider. Major use cases such as exclusions (negation) and plans (value restriction) as well as common information model constructs (for example value and qualified number restrictions) cannot be faithfully represented using DLs with polynomial time subsumption testing [12, 13], and must, at least in runtime be ruled out. However, evidence guiding the choice of representation are still relatively sparse. For example partitioning of ontologies for specific use cases might allow the practical use of logics with non-tractable complexity.

Then, what could be those guiding principles for the choice of representation? From the perspective of Bodenreider et al., the division between the representation of

“biomedical reality” and the representation “how this reality is perceived” (observed) can be understood as the distinction between ontology and epistemology [19]. But the observation of qualities in biomedical reality is in itself a part of biomedical reality, although a part distinct from the quality being observed. So what is from one perspective thought of as belonging to the domain of epistemology may in some other perspective belong to the domain of ontology. Also, as shown by the observables example above, both the ontology of qualities as well as epistemological aspects of qualities (for example observation procedures) may be represented using tractable ontology languages.

If using expressive DLs in runtime for reasoning on typical information models is not possible, there still might be a place for such logics in design time. An alternative could be to encode both ontologies and information models in DLs from which several (computationally tractable) linearisations are generated in order to address different reasoning needs (for example satisfiability checking). Checking validity and conformance to a reference model are examples of design time use cases where satisfiability checking may be applied.

A research program for testing the hypothesis that expressive DL may be used for typical design time EHR representation tasks such as information modelling and terminology binding would include:

- tools for automatically translating *openEHR* archetypes to an OWL representation, including existing terminology

bindings, allowing the testing of different approximations to the archetype semantics,

- translation of the full set of archetypes in the *openEHR* public repository, and
- running typical design-time inferencing tasks.

The demarcation between ontology- and information model representations cannot be easily found in the nature of the entities represented, but rather in the complexity of the kinds of inference needed in a specific use case. As noted by Berzell, the context and the intended use of the EHR will be most important when choosing what methods, principles and technologies to use [20]. This is something that will have to be decided in close cooperation between expertise in both healthcare, ontology and informatics.

Acknowledgements

This work was supported by the DebugIT project of the EU 7th Framework Program grant agreement ICT-2007.5.2-217139.

References

1. ISO. *ISO 12967-2:2009: Health informatics - Service architecture – Part 2: Information viewpoint*. International Organization for Standardization, Geneva, Switzerland, 2009.
2. A.D. Spear. *Ontology for the twenty first century: An introduction with recommendations*. Saarbrücken, Germany: *Institute for Formal Ontology and Medical Information Science*, 2006.
3. Thomas Beale and Sam Heard. An ontology-based model of clinical information. *Studies in Health Technology and Informatics*, 129(Pt 1):760–764, 2007. PMID: 17911819.
4. *openEHR*. The *openEHR* reference model. <http://www.openehr.org/svn/specification/TRUNK/publishing/roadmap.html>, 2010.
5. T. Beale. Archetypes: Constraint-based domain models for future-proof information systems. In *OOPSLA 2002 workshop on behavioural semantics*. Citeseer, 2002.
6. *openEHR*. *openEHR Clinical Knowledge Manager*. <http://www.openehr.org/knowledge/>.
7. Stefan Schulz, Boontawee Suntisrivaraporn, Franz Baader, and Martin Boeker. SNOMED reaching its adolescence: ontologists’ and logicians’ health check. *International Journal of Medical Informatics*, 78 Suppl 1:S86–94, April 2009. PMID: 18789754.
8. HL7 terminfo project. <http://www.hl7.org/special/committees/terminfo/index.cfm>.
9. D. Markwell, L. Sato, and E. Cheetham. Representing clinical information using SNOMED clinical terms with different structural information models. In *KR-MED 2008*, page 72, 2008.
10. C. Martínez-Costa, M. Menárguez-Tortosa, J. T. Fernández-Breis, and J. A. Maldonado. A model-driven approach for representing clinical archetypes for semantic web environments. *Journal of biomedical informatics*, 42(1):150–164, 2009.
11. Roland Hedayat. *Semantic web technologies in the quest for compatible distributed health records*. Master thesis, Uppsala University, 2010.
12. F. Baader, S. Brandt, and C. Lutz. Pushing the EL envelope. In *International Joint Conference on Artificial Intelligence*, 19, page 364, 2005.
13. F. Baader, S. Brandt, and C. Lutz. Pushing the EL envelope further. In *Proceedings of the OWLED 2008 DC Workshop on OWL: Experiences and Directions*, page 22, 2008.
14. O. Kilic, V. Bicer, and A. Dogac. Mapping Archetypes to OWL. *Middle East Technical University, Ankara, Türkiye*, 2006.
15. A.L. Rector and S. Brandt. Why do it the hard way? The case for an expressive description logic for SNOMED. *Journal of the American Medical Informatics Association*, 15(6):744, 2008.
16. Stefan Schulz, Daniel Schober, Christel Daniel, and Marie-Christine Jaulent. Bridging the semantics gap between terminologies, ontologies, and information models. *Studies in Health Technology and Informatics*, 160(Pt 2):1000–1004, 2010. PMID: 20841834.
17. IHTSDO. SNOMED CT® style guide: Observable entities and evaluation procedures (Laboratory), June 2010.
18. C. Mungall, G. Gkoutos, N. Washington, and S. Lewis. Representing phenotypes in OWL. *OWL: Experiences and Directions (OWLED 2007)*, Innsbruck, Austria, 2007.
19. O. Bodenreider, B. Smith, and A. Burgun. The ontology-epistemology divide: A case study in medical terminology. In *Formal ontology in information systems: Proc Third International Conference (FOIS-2004)*, 185–195. IOS Press, 2004.
20. Martin Berzell. *Electronic Healthcare Ontologies: Philosophy, the real world and IT structures*. PhD thesis, Linköping University, 2010.

Ontology-Based Mammography Annotation and Similar Mass Retrieval with SQWRL

Hakan Bulu¹, Adil Alpkocak¹, Pinar Balci²

¹Department of Computer Engineering, Dokuz Eylul University, Izmir, Turkey

²Radiology Department, Dokuz Eylul University Medical School, Izmir, Turkey

This Paper has been withdrawn.

The CHRONIOUS Ontology Suite: Methodology and Design Principles

Luc Schneider, Mathias Brochhausen

Institute for Formal Ontology and Medical Information Science,
Saarland University, Saarbrücken, Germany

Abstract. This paper outlines the methodology and the basic design principles underlying the development of the ontology suite in the EU- funded project CHRONIOUS, comprising a Middle Layer Ontology for Clinical Care (MLOCC) as well as two domain ontologies, one on Chronic Obstructive Pulmonary Disease (COPD) and one on Chronic Kidney Disease (CKD). The article also sketches some major philosophical reflections underpinning the CHRONIOUS ontologies.

Keywords: middle layer ontology, chronic disease ontology, clinical care ontology, chronic obstructive pulmonary disease, chronic kidney disease

1 Overview of the CHRONIOUS Ontology suite

1.1 Purpose

The CHRONIOUS¹ project's primary aim is the development of an integrated telemedical platform for monitoring the general health status of patients with chronic health conditions and providing decision support for the clinicians treating them. For demonstrative purposes, the project focuses on Chronic Obstructive Pulmonary Disease (COPD) and Chronic Kidney Disease (CKD) including Renal Insufficiency [1].

Part of the CHRONIOUS platform is an ontology-powered literature search tool providing efficient and accurate access to recent research literature on COPD and CKD for health care professionals. Publications are annotated both with classes from the CHRONIOUS ontologies and with terms from the MeSH² thesaurus³. Thus, the CHRONIOUS literature search system combines the

terminological knowledge and the multi-linguistic capabilities of MeSH with the clinical expert knowledge encoded by ontologies as topic-neutral representations of the items (objects, processes, qualities, dispositions, functions, etc.) in the domain of the etiology, diagnosis and therapy of COPD and CKD.

1.2 Technical Details

The CHRONIOUS ontologies exhibit the following modular structure:

1. The Middle Layer Ontology for Clinical Care (MLOCC)⁴, which has been extracted from the ACGT Master Ontology [2,3], augmented by general clinical classes identified in the documentary sources supplied by the medical experts in the CHRONIOUS project (see below). MLOCC contains general classes for objects (chemical substances, cells, tissues, organs, technical instruments, etc.), processes, qualities, powers, functions and roles that are relevant to pathological, anatomical, diagnostic and therapeutic aspects of clinical care. MLOCC is based on Basic Formal Ontology [4], a foundational ontology widely used in the biomedical domain; it also contains a subset of the Foundational Model of Anatomy (FMA) [5]

¹ "An Open, Ubiquitous and Adaptive Chronic Disease Management Platform for COPD and Renal Insufficiency" (<http://www.chronious.eu/>)

² Medical Subject Headings; www.nlm.nih.gov/mesh/

³ Initially it was planned to use the very same ontologies also in the decision support system, but because of scalability issues, this idea was not followed up by the project partners.

⁴ <http://www.ifomis.org/chronious/mlocc>

and the Relation Ontology (RO) [6].

2. the COPD ontology⁵ containing specific domain knowledge about Chronic Obstructive Pulmonary Disease, and
3. the CKD ontology⁶ containing specific domain knowledge about chronic kidney diseases and renal insufficiency.

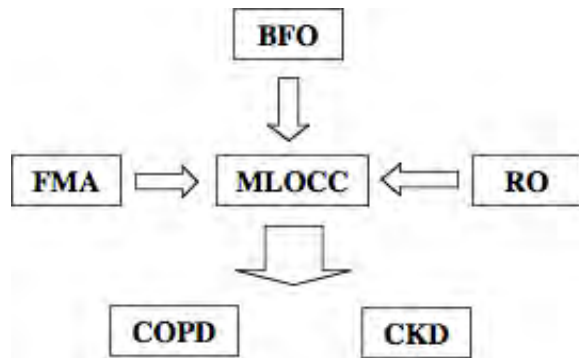


Figure 1. Components of the CHRONIOUS ontology module

The following table sums up some quantitative data about the CHRONIOUS ontologies:

	Classes	Relations	Axioms
MLOCC	476	65	~ 900
COPD (+MLOCC)	964	65	~ 2000
CKD (+MLOCC)	972	65	~ 2000

From the table above the reader can gather that all object properties (relations) used in the CHRONIOUS ontologies (of which the RO relations are a subset) are defined on the level of MLOCC. The COPD and the CKD ontologies have approximately the same size. All ontologies have been developed in OWL-DL using Protégé 4.02 and 4.1; the restriction to OWL-DL may be unnecessary for the purposes of literature search, but (1) the ontologies are intended to be appropriate for other, more reasoning-intensive uses and (2) decidability allows for efficient consistency checking, consistency being a necessary, if not sufficient, condition for adequacy to reality.

2 Methodological Principles

2.1 Realism

From the implementation point of view, an

⁵ <http://www.ifomis.org/chronious/copd>

⁶ <http://www.ifomis.org/chronious/ckd>

ontology is a software artifact (e.g. an OWL file). Hence its design features can be more or less influenced by the peculiar use case(s) of the overall information system in which it is integrated. Thus, to guarantee data exchange and re-use across heterogeneous sources, we must recur to ontologies that unify the different ways in which the domain is viewed by the different end-users and information systems designers. Realist ontologies, i.e. ontologies that aim to depict reality independently of the mental or digital representation of reality by end users and knowledge engineers, provide a unified way of representing the domain from the start, without the need of an ex-post integration of heterogeneous perspectives.

Since the CHRONIOUS literature search system is intended to be user-friendly for health care professionals and clinicians of various specialities, the best option seemed to us to implement the CHRONIOUS ontologies as realist ontologies. Therefore we adopted realism as a fundamental methodological choice, following the general approach already adopted for the development of the ACGT (Advancing Clinico-Genomic Clinical Trials on Cancer) Master Ontology [2, 3].

2.2 Adoption of BFO as a Top-Level Ontology

In order to guarantee that a domain ontology is a reference ontology, it is mandatory to use the best available sources of expert knowledge about the reality to be depicted, e.g. the biomedical domain, but also to recur to formal ontologies, i.e. formal systems of general categories and relations for depicting reality, namely foundational or top-level ontologies such as BFO [4] or DOLCE [7]. Since BFO (Basic Formal Ontology) is used as a reference top-level ontology in the major open-source repository for biomedical ontologies, namely the OBO Foundry [8], we have decided to build the Middle Layer Ontology for Clinical Care (MLOCC) and thus the CHRONIOUS ontologies on re-using BFO by appending the upper-level classes of MLOCC onto leaves of BFO, as already done in the case of the ACGT Master Ontology [2,3]. By appending MLOCC-classes under BFO-classes, the meaning of the latter are given a reality-driven semantics, thus ensuring that the CHRONIOUS ontologies are truly reference ontologies satisfying the

methodological requirement of realism.

Some examples should suffice to illustrate how MLOCC has been constructed around BFO; in the following we assume familiarity with the general structure of BFO [4]. Under the BFO-class *Generically Dependent Continuant* we have appended the MLOCC-class *Information Object*, which covers information artifacts such as questionnaires, medical images or designs (plans, e.g. standardized procedure plans) as distinct from their material supports (paper, traces on electromagnetic storage devices etc.). The BFO-class *Quality* subsumes e.g. the MLOCC-classes *Magnitude* (covering physical magnitudes) and *Condition* (which subsumes the important class *Organismal Condition*). Under the BFO-class *Function* we find the MLOCC-class *Organ Function*, while the BFO-class *Role* covers not only social roles like *Professional Role* or *Administrative Role*, but also biochemical extrinsic features like *Biomarker*, *Drug*, *Catalyst* (Enzyme) or *Hormone*. The BFO-class *Disposition* subsumes the all-important MLOCC-classes *Disease* and *Malfunction*.

Under the BFO-class *Object* we find mainly the MLOCC-nodes *Biological Independent Continuant* (covering *Organism* and *Organismal Independent Continuant* which subsumes classes added from the FMA), *Chemical Substance*, *Institution* and *Technical Object* (for instance devices and instruments). The BFO-class *Process* is subdivided into the MLOCC-classes *Intentional Process* and *Natural Process*; the first subsumes classes related to human and social activities, in particular medical (diagnostic and therapeutic) processes, while the second subsumes *Chemical Process* and *Organismal Process*. An often recurrent feature is the following: a *Disposition* (cf. Figure 2), i.e. a *Disease* or *Malfunction* is realized by an *Organismal Process* that bears the relation *has_Outcome* to a *Quality*, i.e. an *Organismal Condition* or a *Magnitude* such as a *Glomerular Filtration Rate* or *Spirometric Measure*.

2.3 Modularity and Re-use

Another important methodological principle besides realism is perspectivalism, which is the recognition that reality is complex and variegated [4]. There are many different representations of reality that are equally adequate because they capture different

important aspects of the same world. Thus, reality can be assayed in terms of substances and their qualities or powers as well as in terms of processes. More importantly, reality can be described at various levels of granularity, ranging from the atomic and molecular levels to those of cells, tissues and organisms. As a consequence, reality cannot be accounted for in terms of a single monolithic ontology, but only in terms of a multitude of modular ontologies that are orthogonal to each other and thus are also re-usable.



Figure 2. An example of the ontological representation using BFO

We have already mentioned that the CHRONIOUS ontologies were built on top of an established Upper Ontology, namely BFO, as far as high-level classes are concerned. As to relations or object properties, we have recurred to the Relations Ontology (RO) of the OBO Foundry [6]. RO is a set of formal relations that are used in biomedical applications, and minor modifications apart, the object properties of the CHRONIOUS ontologies are an extension of RO. While BFO is directly imported into MLOCC, the object properties in RO have been copied into MLOCC, since the tree of RO object properties has been modified. E.g. in MLOCC *participates_in* subsumes not only *agent_in*, but

also *means_of*.

Another part of MLOCC modelled after an already existing ontology is the branch below the class *Organismal Independent Continuant*, which mirrors the structure and content of the Foundational Model of Anatomy, a reference ontology for anatomy [5]. The FMA classes have been added as is, except for two major exceptions. First, *Biological Macromolecule* has been moved under *Chemical Substance*. Second, the hierarchy below *Cardinal Organ Part* has been simplified by directly adding *Cardinal X Part* for any Organ X (e.g. *Cardinal Heart Part*, *Cardinal Lung Part*). This move was necessary to avoid a ramification of subdivisions that would be spurious for the use case of literature annotation; it does not compromise realism, since the added classes represent objective divisions in biomedical reality.

The principle of re-using already validated software constructs has been applied for the design of the Middle Layer Ontology for Clinical Care (MLOCC) itself, the core of which was extracted from the ACGT Master Ontology [2,3].

Finally, the CHRONIOUS ontologies are modular insofar both the COPD Ontology and the CKD Ontology import MLOCC, which itself imports BFO. MLOCC represents the common core of the chronic disease ontologies, which expand on it in domain-specific ways, except for the object properties that are defined on the level of MLOCC.

Taking modularity and re-use a step further, we would have wished to be able to re-use the Ontology for General Medical Science (OGMS)⁷ as an intermediary level between BFO and MLOCC, but for the fact that this promising medical ontology is still in its early phase of development. Also, we did not consider to use of the Disease Ontology (DO)⁸ because (1) DO is not based on any foundational ontology like BFO, (2) DO does not provide axioms besides trivial subclass axioms and (3) not all subclass-relations in DO are formal ISA-relations resulting occasionally in multiple inheritance of primitive classes (e.g. the class *Nelson's Syndrome* is a subclass to both *Adrenal Cortex Disease* and *Pituitary Neoplasm*).

⁷ <http://code.google.com/p/ogms/>

⁸ http://do-wiki.nubic.northwestern.edu/index.php/Main_Page

2.4 Selection of Sources and Class Extraction

The classes and relations of the CHRONIOUS domain ontologies have been extracted from clinical guidelines about COPD and CKD selected and validated by the medical board of the CHRONIOUS project [9,10,11,12,13,14]. The actual class extraction roughly followed the procedure proposed in [15] and adopted for the ACGT Master Ontology, as indicated in [2]:

1. A glossary of candidate classes is established whose coverage is evaluated by domain experts (e.g. clinicians).
2. The classes are assigned to the different ontological categories, i.e. classes of the foundational ontology BFO.
3. The classes are assigned either to the middle layer (MLOCC) or to the domain ontologies.
4. The classes are ordered in subsumption hierarchies or taxonomies.
5. The (binary) non-taxonomic relations between the classes are identified and represented as object properties.
6. A class dictionary is constituted, describing each class and stating the relations that have it as their domain.
7. The inverse relation and mathematical properties (symmetric, transitive, etc.) of each object property is specified.
8. Formal axioms are stated; these axioms constrain the extension of classes by specifying binary relationships between them in the form of object properties.

3 Design Principles

3.1 General Design Principles

The general design principles discussed below pertain to the articulation of taxonomies; these principles have been explained in detail in [2], so we may review them briefly, demonstrating issues related to applying those principles in the design of the CHRONIOUS ontologies.

Taxonomies should contain only types, not instances or tokens. This principle trivially reflects the type-token distinction. However,

information objects like questionnaires are ontologically tricky in this respect. Is *Chronic Respiratory Disease Questionnaire* a type or a token? We have decided to treat information objects as types and their “copies” as tokens. Nonetheless, not every ontologist may share this view, and consider the relation between an artifact and its copy as being distinct from the relationship between a class and its members. This may seem to be a merely philosophical question, but one should bear in mind that different modeling options exist for information objects and other artifacts that may have a significant impact given the principle mentioned above.

Taxonomies are exclusively based on formal subsumption, i.e. subsumption ties or subclass relationships have to be rigid and context-independent. Time- or situation-dependent taxonomical structures cannot be considered in a (relatively) context-free reference ontology. E.g. when we have to interpret the sentence “Tiotropium bromide is a bronchodilator drug” ontologically, it would seem that a certain amount of tiotropium bromide is used as a bronchodilator: to be a bronchodilator is an extrinsic, not an intrinsic feature of tiotropium bromide. Extrinsic features are roles. Hence we chose to state (in the COPD ontology) that *Tiotropium Bromide* is not a subclass of *Bronchodilator* (and hence *Drug*), and that the latter is not a subclass of *Chemical Substance*. Instead we have classified *Drug* and *Bronchodilator* as *Role*, and have stipulated that *Tiotropium Bromide* (as an *Anticholinergic*) has the role *Bronchodilator*.

The immediate subclasses of a given class should ideally be exhaustive, i.e. their union should cover exactly the whole of the superclass. This principle only half-way cherished in the ACGT Master Ontology was impossible to maintain in the development of the CHRONIOUS ontologies, which are meant to be extendable. Considerations of relevance, as reflected in the sources, also led to the consequence that the children of a node are not exhaustive: this is trivially true e.g. for *Cardinal Heart Part* (MLOCC), *Cardinal Lung Part* (COPD) and *Cardinal Kidney Part* (CKD). Instead of this idealized completeness requirement we have adopted a more pragmatic

completeness requirement, according to which the most relevant subclasses should be present.

Multiple inheritance of primitive classes should be avoided. To return to the bronchodilator example, it was out of the question to subsume *Tiotropium Bromide* both under *Bronchodilator* and *Chemical Substance* (indirectly). Instead of multiple inheritance, one should privilege intrinsic or formal subsumption on the one hand over role attribution on the other (as in the example above).

Primitive sibling classes should be disjoint.

UnknownX as well as other catch-all classes for remaining cases should be avoided. Indeed, it is tempting to render a non-exhaustive subdivision complete by adding a class that covers the rest of the extension of the superclass (e.g. something like *Other Cardinal Lung Part*). However, such a class does not cut at a joint of reality, i.e. it does not represent an ontological, but an epistemological distinction. Indeed, it covers all classes not yet known: yet it is irrelevant for reality whether something is known or not. One needs to distinguish UnknownX-classes from classes like *Undifferentiated Gender* which reflect real borderline cases.

3.2 Specific Design Principles

The following design principles correspond to actual ontological choices based on an intended, philosophical interpretation of Basic Formal Ontology. Note that the intuitions behind the following principles may diverge from those proposed in expositions of BFO such as [4].

Occurrents do not participate in other occurrents. There are no events of events: events are changes and as such do not change. Qualitative change of events or processes is just having different temporal parts at different times. There is no change in the sub-ontology of processes. A process of heart beating does not accelerate, but a heart rate, which is the quality of an organism resulting of a heart beating, does.

Realizable entities (dispositions, functions, roles) do not participate in occurrents. Indeed, realizable entities are expressed or realized by processes; as mere potentialities they do not change per se.

This means that the renal filtration function does not change, but is realized by the process of renal filtration which has a glomerular filtration rate (GFR) as an outcome. The GFR may increase or decrease, the renal filtration function not. In general, everyday talk about increase or decrease of functions translates into the increase or decrease of qualities that result or are affected by processes realizing dispositions, functions or roles.

Realizable entities only characterize independent continuants, except roles. The latter exception is necessary if we want to exploit role assignment in order to avoid non-rigid subsumption or multiple inheritance. E.g. symptoms are roles of organismal processes; they cannot be intrinsic features, since an organismal process need not be a symptom of a specific disease irrespective of other co-occurrent events.

4 Conclusion

Our aim was to provide a general overview regarding the methodology and the design principles leading the implementation of the CHRONIOUS ontologies, as well as an idea of their overall structure. We stressed the importance of the methodological criteria of realism and modularity and we have sketched the philosophical positions guiding the construction of these ontologies. We have shown that the CHRONIOUS ontologies represent a carefully thought-through and expert-validated contribution to (a) the general ontology of (chronic) diseases and pathological conditions as well as to (b) the ontology of specific chronic illnesses like COPD and CKD. Ongoing work on MLOCC includes its integration with the the Ontology for General Medical Science (OGMS) as well as the Information Artifact Ontology⁹.

Acknowledgement

Research leading up to the present article has been supported by the ICT-2007-1-216461 grant within the Seventh Framework Programme of the EU, as well as by a post-doc grant from the National Research Fund, Luxembourg (cofunded under the Marie Curie

Actions of the European Commission [FP7-COFUND]), and has been carried out under subcontract to the Fraunhofer Institute for Biomedical Engineering, St. Ingbert (Germany).

References

1. Farré F, Papadopoulos A, Munaro G, Rosso R. An Open, Ubiquitous and Adaptive Chronic Disease Management Platform for Chronic Respiratory and Renal Diseases (CHRONIOUS). In: Conley, EC, Doarn, C, Hajjam-El-Hassani, A, editors. Proceedings of the International Conference on eHealth, Telemedicine, and Social Medicine; 2009 Feb 1-7; Cancun, Mexico. New York: IEEE Press; 2009. p.184-9.
2. Cocos, C. Design Principles of the ACGT Master Ontology: Examples and Discussion. Saarbrücken: Institute for Formal Ontology and Medical Information Science; 2008 [cited 2011 Jun 5]. Available from: [http://www.ifomis.org/wiki/ACGT_Master_Ontology_\(MO\)](http://www.ifomis.org/wiki/ACGT_Master_Ontology_(MO))
3. Brochhausen M, Spear AD, Cocos C, Weiler G, Martin L, Anguita A, et al. The ACGT Master Ontology and Its Applications – Towards an Ontology-Driven Cancer Research and Management System. J Biomed Inform. 2011; 44: 8–25.
4. Spear, A. Ontology for the Twenty First Century: An Introduction with Recommendations. Saarbrücken: Institute for Formal Ontology and Medical Information Science; 2006. Available from: <http://www.ifomis.org/bfo/documents/manual.pdf>
5. Rosse C, Mejino JVL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003; 36: 478-500.
6. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in Biomedical Ontologies. Genome Biol. 2006; 6: R46.
7. Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening Ontologies with DOLCE. In: Gomez-Perez A, Benjamins VR, editors. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Proceedings of the 13th International Conference, EKAW 2002; 2002 Oct 1-4; Sigüenza, Spain; Heidelberg: Springer; 2003. P. 166-81
8. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5
9. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the

⁹ http://www.obofoundry.org/cgi-bin/detail.cgi?id=information_artifact

- diagnosis, management, and prevention of chronic obstructive pulmonary disease. Bethesda (MD): Global Initiative for Chronic Obstructive Lung Disease (GOLD); 2009.
10. National Kidney Foundation Kidney Disease Outcomes Quality Initiative (NKF KDOQI). KDOQI Clinical Practice Guidelines on Hypertension and Antihypertensive Agents in Chronic Kidney Disease. New York (NY): National Kidney Foundation; 2004 [cited 2011 Jun 5]. Available from: http://www.kidney.org/professionals/KDOQI/guidelines_bp/index.htm.
 11. National Kidney Foundation Kidney Disease Outcomes Quality Initiative (NKF KDOQI). KDOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification. New York (NY): National Kidney Foundation; 2002. Available from: http://www.kidney.org/professionals/KDOQI/guidelines_ckd/toc.htm
 12. National Kidney Foundation Kidney Disease Outcomes Quality Initiative (NKF KDOQI). KDOQI Clinical Practice Guidelines for Bone Metabolism and Disease in Chronic Kidney Disease. New York (NY): National Kidney Foundation; 2003. Available from: http://www.kidney.org/professionals/KDOQI/guidelines_bone/index.htm
 13. National Kidney Foundation Kidney Disease Outcomes Quality Initiative (NKF KDOQI). KDOQI Clinical Practice Guidelines for Managing Dyslipidemias in Chronic Kidney Disease. New York (NY): National Kidney Foundation; 2003 [cited 2011 Jun 5]. Available from: http://www.kidney.org/professionals/KDOQI/guidelines_lipids/toc.htm
 14. National Kidney Foundation Kidney Disease Outcomes Quality Initiative (NKF KDOQI). KDOQI Clinical Practice Guidelines and Clinical Practice Recommendations for Diabetes and Chronic Kidney Disease. New York (NY): National Kidney Foundation, 2007: http://www.kidney.org/professionals/kdoqi/guidelines_anemia/references.htm
 15. Gomez-Perez A, Fernandez-Lopez M, Corcho O. Ontological Engineering. London: Springer; 2004.

Applying Rigidity to Standardizing OBO Foundry Candidate Ontologies

A. Patrice Seyed, Stuart C. Shapiro

Department of Computer Science and Engineering, University at Buffalo, State University of New York, USA

Abstract. The Open Biomedical Ontology (OBO) Foundry initiative is a collaborative effort for developing interoperable, science-based ontologies. OBO uses the Basic Formal Ontology (BFO) as its upper ontology. Ontologies developed for OBO use include some that have been ratified, and others holding the status of candidate. There are no formal, principled criteria that a candidate ontology must meet for ratification. To help address this problem, we propose a formal integration between Rigidity, a major component of OntoClean's approach to quality assurance of ontologies, and BFO's theory of types. This work augments ongoing efforts to build software designed to evaluate and standardize OBO Foundry candidate ontologies.

Keywords: ontology, criteria, OBO Foundry, BFO, OntoClean

1 Introduction

The Open Biomedical Ontology (OBO) Foundry¹ initiative is a collaborative effort for developing interoperable, science-based ontologies. A recently adopted principle for these ontologies is that they use the Basic Formal Ontology (BFO) [1] as their upper ontology. Some OBO Foundry ontologies have been ratified, and others hold the status of candidate. Rigidity is a major component of the OntoClean approach for detecting when the taxonomic relation is being used improperly [2]. A property is Rigid, if it is essential to all its instances; Non-Rigid, if non-essential to some instance; or Anti-Rigid, if non-essential to all instances.

BFO is only partially logically axiomatized [1]. Domain experts developing OBO Foundry candidate ontologies must regularly query BFO-trained ontologists in order to adhere to BFO's principles. Currently, there are no formal, principled criteria that a candidate ontology must meet for ratification. To address this problem, we propose a formal integration between OntoClean's theory of Rigidity and BFO's theory of *types*. We also propose an approach for evaluating OBO Foundry candidate ontologies based on this integration.

2 Formal Theory of Classes

OntoClean uses *properties* as its categorical unit, which are the intension, or meaning, of general terms. BFO uses *types*, which are defined as that in reality to which the general terms of science refer (B. Smith, personal communication). We unify property and types under *class*. In what follows, we assume a first-order, sorted logic.

Although there are many theories of existence, we introduce a relation, ***exists_at*** (x, t), which is non-committal and means that, under a certain ontological theory, object x is within its domain and x 's existence spans some time, t . Everything exists at some time:

Axiom 1. $\forall x \exists t(\mathbf{exists_at}(x, t))$

member_of(x, A, t) means that object x satisfies the definition of class A at t . With the ***member_of***(x, A, t) relation, there is no commitment about the nature of A . Therefore, membership at a time *does not* presuppose that existence spans that time:

Axiom 2. $\neg \forall xt(\exists A \mathbf{member_of}(x, A, t))$
 $\rightarrow \mathbf{exists_at}(x, t)$

A particular class might or might not satisfy the unary predicate ***Instantiated***, which means there is some member of A at t

¹ <http://www.obofoundry.org>

that exists at t :

Definition 1. *Instantiated*(A) =_{def}

$$\exists xt(\text{member_of}(x,A,t) \wedge \text{exists_at}(x,t))$$

If a class does not have any members at any time, it satisfies the predicate **Empty**:

Definition 2. *Empty*(A) =_{def}

$$\neg \exists xt(\text{member_of}(x,A,t))$$

Empty (*Full_Eye_Transplant*) holds because no such procedure has been performed yet.

If a class has as members only those objects that exist at all times at which they are members, it satisfies the predicate **Members_Exist**:

Definition 3. *Members_Exist*(A) =_{def}

$$\forall xt(\text{member_of}(x,A,t) \rightarrow \text{exists_at}(x,t))$$

Assuming a class *Animal* is defined to have as members animals only at times they are alive, **Members_Exist** (*Animal*) holds.²

3 Reformulating Rigidity

Rigidity has been defined in terms of S5 modal logic. As part of our integration, we provide just the underlying intuitions of those modal formalisms, prior to reformulating Rigidity in our formal system. Each object that has a Rigid property has that property at all times at which the object exists. We formalize this in terms of classes, instead of properties, by the predicate **Rigid**:

Definition 4. *Rigid*(A) =_{def}

$$\forall x(\exists t(\text{member_of}(x,A,t))$$

$$\rightarrow \forall t_1(\text{exists_at}(x,t_1) \rightarrow \text{member_of}(x,A,t_1)))$$

Rigid (*Person*) means that all members of the class *Person* are people at all times at which they exist.

As an amendment to the original formulation of Rigid, [3] proposes that Rigid properties are only instantiated by actually existing objects. We have captured this intuition separately from Rigid, under the **Members_Exist** predicate. Also, because unexemplifiable properties are trivially Rigid, [3] constrains the theory (as suggested by [4]

and [5]) to properties for which there exists some instance. We have separately defined this notion, also, under the **Instantiated** predicate.

Non-Rigid is the negation of Rigid, which we apply for our class formulation under the predicate **Non-Rigid**:

Definition 5. *Non-Rigid*(A) =_{def} \neg *Rigid*(A)

Assuming that a person is a member of Student only while a registered student, **Non-Rigid** (*Student*) holds.

Anti-Rigid is true of a property, if, for every object that has that property, it is *possible* that it does not have that property at some time. An object may have an Anti-Rigid property at all times at which it exists. BFO is not concerned with what could have been, but rather what has been or currently is; therefore, Anti-Rigid is irrelevant to our theory.

4 Integrating Rigidity with BFO Theory of Types

The objects of BFO's domain are partitioned into *particulars* and *types*. Particulars are entities confined to specific spatial, spatiotemporal, or temporal regions (e.g., a specific grasshopper in front of me, its life, or the time interval that its life spans, respectively). Under BFO's theory, existence of a particular is based on it being observable at some level of granularity and/or causal by some scientifically-based measure. Numbers, for example, do not exist in BFO. **Type** (A) means that A is a class that meets the criteria for being a type, which we provide in what follows.

Not all classes thought to be types satisfy our reformulation of Rigid, for example *Embryo* and *Fetus*. If an organism maintains its identity through its development from an embryo into a fetus, then both classes are Non-Rigid. If these classes are not in fact types, then our axiomatization is clear. However, whether these sorts of classes are types or not is still debated by the OBO Foundry community; therefore, this issue remains unresolved. For the purposes of our method, we exclude these controversial classes from our domain; hence, types satisfy **Rigid**:

Axiom 3. $\forall A(\text{Type}(A) \rightarrow \text{Rigid}(A))$

² For organisms we equate existence with living.

Types must also be instantiated [6]:

Axiom 4. $\forall A(\text{Type}(A) \rightarrow \text{Instantiated}(A))$

Types are therefore non-empty:³

Theorem 1. $\forall A(\text{Type}(A) \rightarrow \neg \text{Empty}(A))$

Another criterion for every class that is a type is that every member of the class at a time exists at that time. Therefore, every type satisfies *Members_Exist*:

Axiom 5. $\forall A(\text{Type}(A) \rightarrow \text{Members_Exist}(A))$

instance_of(x, A, t) means that particular x is an instance of type A at time t . If a general term refers to a class that is a BFO type, then each of the members of the class instantiates the type:

Definition 6. $\text{instance_of}(x, A, t) =_{\text{def}} \text{member_of}(x, A, t) \wedge \text{Type}(A)$

While there is no restriction on what objects can be members of a class, particulars, not types, are instances of a type:

Axiom 6. $\forall AB(\text{Type}(A) \wedge \text{Type}(B) \rightarrow \neg \exists t(\text{instance_of}(A, B, t)))$

A class which satisfies *Instantiated* but not *Members_Exist* satisfies the predicate *Partial*:

Definition 7. $\text{Partial}(A) =_{\text{def}} \text{Instantiated}(A) \wedge \neg \text{Members_Exist}(A)$

isa (A, B) means that all instances of type A are instances of type B :

Definition 8. $\text{isa}(A, B) =_{\text{def}} \forall xt(\text{instance_of}(x, A, t) \rightarrow \text{instance_of}(x, B, t))$

isa is a relation between types:

Axiom 7. $\forall AB(\text{isa}(A, B) \rightarrow \text{Type}(A) \wedge \text{Type}(B))$

It is the “backbone” BFO relation for scientific

classification, i.e., building taxonomies. *isa* is provably reflexive, transitive, and anti-symmetric.

Under OntoClean’s modal formulations, no Anti-Rigid property is a parent of a Rigid property (although a Rigid property may have a Non-Rigid parent, and vice versa). Although Anti-Rigid is irrelevant to our theory, by our reformulation of Non-Rigid, no *Non-Rigid* class is part of an *isa* hierarchy:

Theorem 2. $\forall A(\text{Non-Rigid}(A) \rightarrow \forall B(\neg \text{isa}(A, B) \wedge \neg \text{isa}(B, A)))$

disa (A, B) (‘d’ for “direct”) means there is no other type “in between” A and B in the *isa* hierarchy:

Definition 9. $\text{disa}(A, B) =_{\text{def}} \text{isa}(A, B) \wedge A \neq B \wedge \forall C(\text{isa}(A, C) \wedge \text{isa}(C, B) \rightarrow C=A \vee C=B)$

disa is provably irreflexive, intransitive, and asymmetric, and *isa* is its transitive closure.

The root type of the BFO upper ontology is *Entity*; *Continuant* and *Occurrent* are its subtypes. Continuants (e.g., a heart) can exist fully at different time instants, while occurrents (e.g., the process of a heart beating) unfold over time.

Following Aristotle’s division of objects into substances and accidents, the two subtypes of *Continuant* are *IndependentContinuant* (IC) and *DependentContinuant* (DC), respectively. (For reasons of space, we omit treatment of the DC subtype *GenericallyDependentContinuant* (GDC), and restrict our discussion to the DC subtype *SpecificallyDependentContinuant* (SDC).) The shape of a specific cell instantiates SDC, and “depends on” a specific cell, which instantiates IC. *depends_on* (x, y, t) means that the specifically dependent continuant x exists at t only if the independent continuant y exists at t :⁴

Axiom 8. $\forall xy(\exists t(\text{depends_on}(x, y, t)) \rightarrow \forall t_1(\text{exists_at}(x, t_1) \rightarrow \text{exists_at}(y, t_1)))$

It also means that x cannot migrate to another independent continuant:

³ Informal proofs corresponding to the theorems presented here are provided at <http://www.cse.buffalo.edu/~apseyed/icbo2011proofs.pdf>.

⁴ This relation is frequently given as ‘inheres’ in the BFO literature.

Axiom 9. $\forall xy(\exists tt_1 \text{ depends_on}(x,y,t) \wedge \text{ depends_on}(x,z,t_1) \rightarrow y=z)$

Depends_On (A,B) means that for every instance of A there is some instance of B where the former instance depends on the latter:

Definition 10. $\text{Depends_On}(A,B) =_{\text{def}} \forall xt(\text{instance_of}(x,A,t) \rightarrow$

$\exists y(\text{instance_of}(y,B,t) \wedge \text{ depends_on}(x,y,t))$

If we assume the class *Student* has as members people at times at which they have the role of student, *Student* satisfies *Non-Rigid* and is not a type. However if the class is “re-conceived” as having as members individual student roles that are instances of *SDC*, then the class does satisfy *Type* and **Depends_On** (*Student*, *Person*) holds. At each time t at which some person x is a student, there exists some y that is a student role and is dependent on x :

$\exists y(\text{instance_of}(y,\text{Student},t) \wedge \text{ depends_on}(x,y,t)).$

BFO’s theory of types is also committed to the *Disjointness Principle*,⁵ that two types have no instances in common unless one is a subtype of the other:

Axiom 10. $\forall AB(\exists xt(\text{instance_of}(x,A,t) \wedge \text{ instance_of}(x,B,t)) \rightarrow \text{ isa}(A,B) \vee \text{ isa}(B,A))$

The *Single Inheritance Principle* follows, that no type has more than one direct supertype:

Theorem 3. $\forall AB(\text{ disa}(A,B) \rightarrow \forall C(\text{ disa}(A,C) \leftrightarrow C=B))$

A version of this principle is advocated by [7] for primitive class hierarchies, in order to keep ontologies modular. The *Disjointness*

⁵ Our work is based on BFO version 1.1, which we consider stable and “frozen” for our research. Recent work [8] indicates this principle only applies to the asserted isa hierarchy. This topic remains under debate.

Principle assists in maintaining the ontological partitioning of types into *DC*, *IC*, and *Occurrent*. Candidates (i.e., classes proposed as types in an OBO Foundry candidate ontology) conceived such that they that violate the *Disjointness* or *Single Inheritance* principles do not satisfy **Type**.

We propose that the subtyping relation between upper ontology types is **disa**, based on the assumption that the types of BFO’s upper ontology fall within a finite domain. If additional types are added, then it is a different ontology.

We also define a relation **disjoint_from** which holds between types A and B iff A and B do not share any instances at any time:

Definition 11. $\text{disjoint_from}(A,B) =_{\text{def}} \forall xt(\text{instance_of}(x,A,t) \rightarrow \neg \text{ instance_of}(x,B,t))$

Axiom 11. $\forall AB(\text{ disjoint_from}(A,B) \rightarrow \text{ Type}(A) \wedge \text{ Type}(B))$

We can show that for two direct subtypes of a third type, if the two types are not identical, then they are disjoint:

Theorem 4. $\forall AB((\exists C(\text{ disa}(A,C) \wedge \text{ disa}(B,C)) \wedge A \neq B) \rightarrow \text{ disjoint_from}(A,B))$

We can also prove that sibling BFO upper ontology types (e.g., *Continuant* and *Occurrent*) and, more generally, any types not related by **isa**, are disjoint types:

Theorem 5. $\forall AB((\text{ Type}(A) \wedge \text{ Type}(B)) \rightarrow (\text{ isa}(A,B) \vee \text{ isa}(B,A)) \oplus \text{ disjoint_from}(A,B))$

5 Applying Rigidity and Other Type Criteria to Standardizing Candidate Types

Isolating violations of the *Disjointness Principle* will assist a modeler in determining if their candidates are types. These violations follow the pattern:

$$\text{ isa}(A,B) \wedge \text{ isa}(A,C)$$

where it does not hold that:

$$\textit{isa}(B,C) \vee \textit{isa}(C,B)$$

which can be inferred under closed-world reasoning, or, it may be that the negation of both disjuncts holds. The potential ontology changes that alleviate this violation include:

1. *isa* (*B,C*) or *isa* (*C,B*) holds.
2. *isa* (*A,B*) or *isa* (*A,C*) is removed, including the choice that *isa* is changed to another relation, e.g., *Depends_On*.
3. *A* is partitioned into multiple candidates, some of which are subtypes of *B* and some of *C*.

One reason for solution #1 is that one (or both) of the disjuncts holds, but the disjunct(s) has not been specified yet by the modeler. A common reason for solution #2 is that one candidate, *B*, is a type, and the other, *C*, is a **Non-Rigid** class. #3 is appropriate if a candidate is evaluated to have as members instances of disjoint upper ontology types (which we term **Heterogeneous**).

We aim to assist a modeler in creating an ontology that does not violate the Disjointness Principle, by preemptively addressing the modeling choices #1, #2, and #3 above. We present a decision tree (see Figure 1)⁶ that assists a modeler in evaluating whether a candidate is a type according to criteria provided in the previous section (satisfying **Instantiated**, **Members_Exist**, and **Rigid**) and, if not, assists in redefining the candidate such that it is consistent with BFO. We assume that a modeler presents her candidates one at a time to a procedure that uses the decision tree to classify each in turn. A candidate that satisfies any combination of **Empty**, **Partial**, **Heterogeneous**, or **¬Members_Exist** satisfies **¬Type** and requires further inspection and re-conceptualization for it to satisfy **Type**.

In Figure 1, the descriptions of the answer choices for Question 2 correspond to more commonly modeled types under *IC*, *DC*, and *Occurrent*, namely *MaterialEntity*, *SDC*, and *Process*. This approach excludes rarely modeled types from our evaluation work, such as

SpatialRegion, *TemporalRegion*, and *SpatioTemporalRegion*.⁷ There are certain other types, (e.g., *GDC*) that will appear in an expanded version of the tree, in future work. The classification of candidates under domain-level types already classified via applications of the decision tree will also be addressed in future work.

5.1 Use Case

Figure 2 shows two candidates, their assumed definitions, and a modeler's response to each question presented to her. The candidate's class label is used within Question 1; for example, for Candidate 1, the question asked is "What is an example of a compound?" Because Question 7 is reached and answered by "no", *Compound* is a type under our approach. Because of the answer "a" given for Question 2, *Compound* is classified under BFO's *MaterialEntity* type.

For Candidate 2, because the answer given for Question 6 is "yes", *Reactant* satisfies **¬Type**, because it satisfies **Non-Rigid**. Question 8 attempts to confirm if the modeler's class definition implicitly refers to some specifically dependent continuant. If this question is answered with "yes", then the candidate is a type if re-conceived as a subtype of *SDC*. Question 9 is asked to determine the classification of the members of *A* (as it was originally conceived) under *MaterialEntity*. In this case, the modeler chose *Compound*, and as a result **Depends_on** (*Reactant*, *Compound*) is asserted. Figure 3 shows the resulting ontology portion, where the upper ontology types are shaded.

Our use case addresses the modeling of Rigid and Non-Rigid candidates, and how a Non-Rigid candidate can be redefined such that it is consistent with BFO, proactively preventing violations of the Disjointness Principle. To extend our use case, if a third candidate, *Sodium Chloride*, were introduced, then the modeler would provide the same answers as given for *Compound*, and *Sodium Chloride* will be subtyped under *MaterialEntity* (the modeler will be able to

⁶ Redundant subtrees for Question 2 choices a, b, or c are combined. Variables that represent modeler-input terms appear in square brackets.

⁷ That the modeling of these types is rare is apparent in the OBO Foundry's Ontology for Biomedical Investigations (see <http://purl.obo.library.org/obo/obi.owl>).

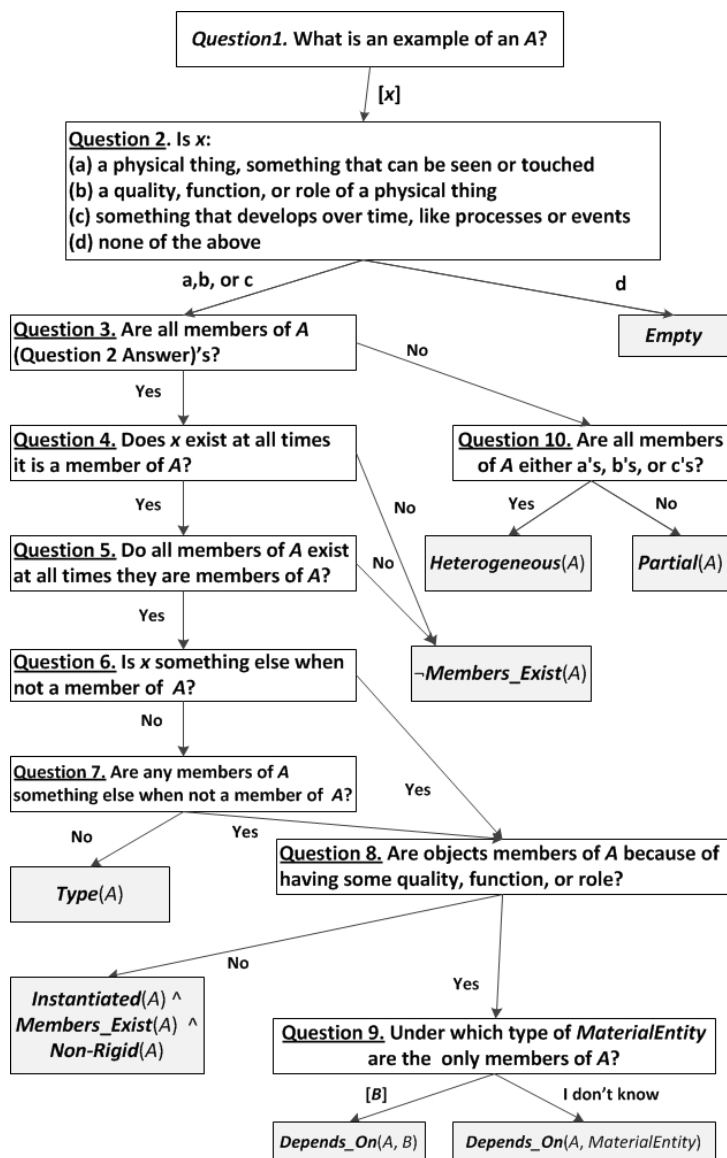


Figure 1. Decision Tree for Standardizing a Candidate Type

Question	Candidate 1: <i>Compound</i>	Candidate 2: <i>Reactant</i>
	Assumed Definition: "A substance consisting of two or more different elements combined in a fixed ratio" [10].	Assumed Definition: "The electron donor in a redox reaction" [10].
1	"sodium chloride in this container"	"sodium chloride in this container"
2	a	a
3	yes	yes
4	yes	yes
5	yes	yes
6	no	yes
7	no	-
8	-	yes
9	-	<i>Compound</i>
10	-	-

Figure 2. Responses to Questions in Decision Tree

subtype *Sodium Chloride* under *Compound* in a future version of the decision tree). Subtyping *Sodium Chloride* under the *SDC* subtree, where *Reactant* is subtyped, will simply not be presented as an option.

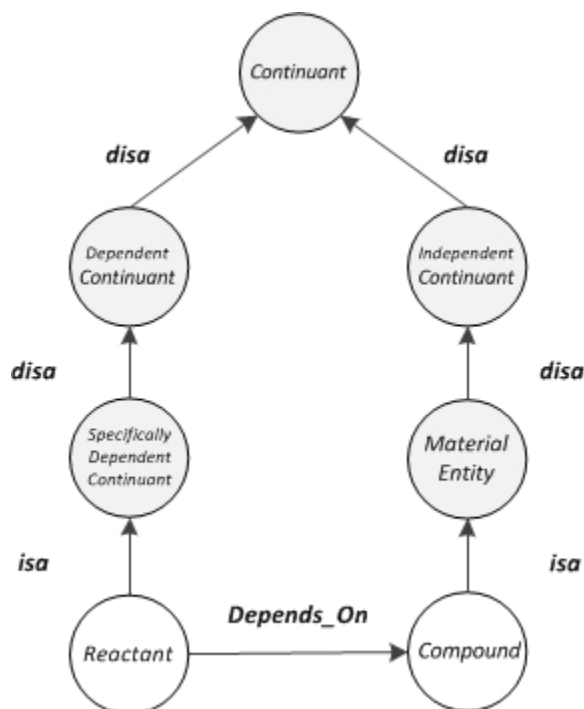


Figure 3. Ontology Portion After Evaluation

6 Conclusion and Future Work

The notion of class covers both OntoClean's notion of property and BFO's notion of *type*. A class might or might not satisfy ***Instantiated***, ***Empty***, ***Heterogeneous***, ***Partial***, ***Members Exist***, ***Rigid***, or ***Non-Rigid***, the latter two capturing the intuitions of Rigidity within our formal theory of classes. BFO's notion of type is captured by a class that satisfies ***Instantiated***, ***Members Exist***, and ***Rigid***. A domain modeler who wants her ontology to be ratified for OBO use and thus BFO-compliant must show that the candidate types of her ontology are indeed types by these criteria.

In the future, we will address whether the Disjointness Principle should be enforced for only what is considered the “primitive” backbone of an ontology. We will also expand the decision tree to address Non-Rigid classes that refer to some material entity (e.g., *Endocrine System*) or process (e.g., *Fertilization*), where their members are conceived as the parts or participants, respectively.

Acknowledgments

We would like to thank William J. Rapaport, Alan Ruttenberg, Barry Smith, and the anonymous reviewers for their comments on previous drafts.

References

1. Spear, A.: *Ontology for the Twenty First Century* (2007)
2. Welty, C. and Guarino N.: Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering*. vol. 39, no. 1, 51–74, (2001)
3. Welty, C. and Andersen. W.: Towards OntoClean 2.0: A framework for Rigidity. *Applied Ontology*. vol. 1, no. 1, 107–116, IOS Press, Amsterdam (2005)
4. Andersen, W. and Menzel, C.: Modal rigidity in the OntoClean methodology. *FOIS* (2004)
5. Carrara, M.: Identity and Modality in OntoClean. *Applied Ontology*. no. 1, vol. 1., 128–139 (2004)
6. Smith, B.: The Logic of Biological Classification and the Foundations of Biomedical Ontology. *International Conference on Logic, Methodology and Philosophy of Science*. Elsevier-North-Holland (2003)
7. Rector, A.: Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. *KCAP* (2003)
8. Smith, B. and Ceusters W.: Ontological realism: A Methodology for coordinated evolution of scientific ontologies. *Applied Ontology*. vol. 5, no. 3, 139–188 IOS Press (2010)
9. Campbell N. A., Reece, J. B.: *Biology*, 8th Edition. Pearson, University of Chicago press (1990)

Higgs Bosons, Mars Missions, and Unicorn Delusions: How to Deal with Terms of Dubious Reference in Scientific Ontologies

Stefan Schulz^{1,2}, Mathias Brochhausen³, Robert Hoehndorf⁴

¹Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

²Institute for Medical Biometry and Medical Informatics, Freiburg University Medical Center, Germany

³Institute for Formal Ontology and Medical Information Science, Saarland University, Saarbrücken, Germany

⁴Department of Genetics, University of Cambridge, UK

Abstract. Realist ontologies claim to represent what exists. Scientific discourse, however, often contains terms of dubious reference when describing current or past hypotheses, plans, or ideas. We present a framework in which a realist ontology is embedded in a description logics theory, the latter being indifferent regarding the existence of class members. It therefore may include units for various kinds of such terms. Using a taxonomy of terminological units we are able to distinguish between different kinds of classes in description logics theories based on whether the classes are believed to have instances or not. We also demonstrate how discourse using terms of dubious reference can be represented without departing from the principle of realist ontologies. An example OWL file can be downloaded from: <http://purl.org/steschu/misc/ICBO2011>.

1 Introduction

Biomedical ontologies are increasingly advocated for the representation of scientific discourse [1], available both as human language content in free-text narratives and the structured content in scientific databases. For this purpose considerable effort has been put into a steadily growing repository of ontologies in biology and medicine, made available through the NCBO BioPortal [2] and the OBO Foundry [3,4].

Upper level ontologies like the Basic Formal Ontology (BFO) [5] or DOLCE [6] ontology, or domain specific top levels like BioTop [7], GFO-Bio [8] and the OBO Relation ontology [9] have been proposed as ordering and constraining principles guiding the development of ontologies in the field of biomedicine.

The OBO Foundry [4] is a collection of biomedical ontologies, following explicit guidelines and adopting a systematic approach to ontology building which is grounded in theoretically well-founded upper-level ontologies and quality criteria. The OBO Foundry recommends BFO as upper level ontology, which adopts a position of *ontological realism*.

In contrast to the proposal that biomedical terminologies should be concept-oriented [10], ontological realism claims that classes in an ontology extend *universals*¹ that exist in reality [12]. Universals, often equated with types, are considered to be entities that exist in their instances. Thus, there can be no uninstantiated universal.

A representational framework increasingly used for biomedical ontologies is Description Logics (DL) [13]. In particular the DL dialects which are standardized by the Semantic Web community as the Web Ontology Language (OWL) [14] have no means to differentiate between universals and non-universals, or between classes that have members in the world and classes that haven't. For example, we can create the classes *Animal* and *Plant* intended to correspond to the universals *Animal*, and *Plant*, respectively. DL theories tolerate nonsense classes like *Planimal* (the conjunction of *Animal* and *Plant*), which have no instances at all. We can state that *Animal* and *Plant* have no instances in common by

¹ As a less restrictive alternative to universals, the more general term *repeatables* has also been recommended [11]. For the sake of simplicity we here use the term "universal" in a broader sense.

declaring them as disjoint. Here, the DL axioms ‘*Animal* and *Plant* **SubClassOf** *Nothing*’ and ‘*Planimal* **SubClassOf** *Nothing*’ are equivalent.

DL theories are *ontologically neutral*: there is no impediment to creating classes that are not the extensions of universals, or to creating classes that have no instances. Whereas the latter is contrary to the basic assumption of realist ontologies, the former construct, so-called defined classes [15], e.g. ‘*Person born in Belgium*’ have been accepted in principle for inclusion in OBO Foundry ontologies [16].

Scientific discourse and discourse in clinical practice often contains terms for which the existence of instances in this world is hypothetical, unknown, purely fictional, or scientifically implausible. End users and application builders may want to use some kind of semantic artefact to link these terms to, for example in the semantic annotation of natural language text.

Ontologists who defend a realist view are pressed to find ways to properly represent assertions using terms for which it is believed or known that no instances exist, without abandoning their principles. In this paper, we investigate how this goal can be achieved. We will use the name *terms of dubious reference* (TDRs) for language expressions which do not denote anything in reality.

We will outline a formal approach of how TDRs are semantically accounted for in description logics based theories in a way which as little as possible contradicts the principles of realist ontologies. To this end we will sketch a typology of TDRs which are commonly used in scientific and clinical discourse.

2 Typology of Terms of Dubious Reference (TDRs)

We define as TDRs those expressions in a natural or logic language, for which the reference to something that exists or has existed is ruled out or hypothetical. We distinguish (see Table 1 for examples):

1. Speculative entities, whose existence cannot be derived from accepted scientific theories, but which are referred to in the discourse in several branches of medicine, e.g. in family practice or complementary medicine.

2. Associations of signs or symptoms which are referred to as a unity according to culture-specific constructs about health and disease, although not backed by any scientific theory.
3. Hypothetical entities, whose existence is postulated by accepted scientific theories, although there is no evidence for their existence. Suitable experiments are expected to collect this evidence.
4. Artefacts or processes that may come into existence only after the execution of a plan in the future. This also includes non-existent chemical compounds, whose structures do not offend the rules of valence but which have never been prepared or found in nature.
5. Fictional concepts, whose existence is not even speculatively assumed but which may be used as metaphors or imaginary patterns.
6. Historic concepts with consensus about their obsolescence.

We exclude from this typology shortcut interpretations suggested by naively interpreted syntactic compositions (missing thumbs, prevented pregnancies, ...), as well as terms including negative statements (non-insulin-dependent diabetes mellitus), because in all those cases there can be referents in reality once we avoid interpreting these expressions at face value.

3 TDRs in Biomedical Ontologies

In the following we will propose a formal solution as to how TDRs can be accommodated in realist ontologies based on classes in description logics. We distinguish between two approaches. In the first approach, the information that a class is not the extension of a universal is not made explicit. In the second approach, it is conveyed by an additional ontology layer.

Throughout this paper we clearly distinguish between *realist ontologies* and *DL models*. Realist ontologies may be – at least partly – represented as DL models, whereas DL models may include units which are irrelevant for realist ontologies.

Category of Entity / Concept of Reference	Term	Definition	Occurrences in MEDLINE
Speculative, not scientifically plausible	Qi	The vital life force in the body, supposedly able to be regulated by acupuncture	299
Hypothetical	Higg's Boson	Hypothetical elementary particle predicted to exist by the standard model of particle physics.	19
Culture-specific constructs	Koro	Syndrome of someone's belief that his/her external genitals disappear	139
Subjects of unrealized plans	Manned mars mission	Process during which a human disembarks on Mars	8
Fictional concepts	Unicorn	Horse with a horn	65
Obsolete concepts of historic interest	Phlogiston	a fire-like element released during combustion	18

Table 1. Typology for NRUs and their occurrences in the MEDLINE abstracts

3.1 Ontology Without Explicit Reference to Universals

The main tenet is the following: While some classes in a DL theory may correspond to universals, there are other classes that do not. In the subsequent formalisms we will use Underline for universals, *Italics* with leading capital for classes extending universals and *lower case italics* for classes not extending universals. **Bold face** is used for individuals and relations involving individuals, for all other relations *lower case italics* are used. For DL expressions we use the OWL Manchester syntax [17].

The extension function *ext* relates a universal to the class of instances of this universal, e.g.

$$J \text{ EquivalentTo } ext(J) \quad (1)$$

$$K \text{ EquivalentTo } ext(K) \quad (2)$$

Using DL syntax we can easily generate new classes that are not the extensions of universals, assuming that the set of universals is not closed under DL operators:

$$b \text{ EquivalentTo } J \text{ and rel some } K \quad (3)$$

However, the DL syntax does not include special symbols for universals or the extension relation *ext*. Formula (1) and (2) are therefore not made explicit, and we therefore refrain from an explicit reference to universals. The following definition, based on the classes *MarsMission* and *Human* as well as the relation **hasParticipant**, does not make any claim on the existence of instances of *mannedMarsMission*:

$$mannedMarsMission \text{ EquivalentTo } MarsMission \text{ and hasParticipant some } Human$$

$$Human \quad (4)$$

This class has no instances in our world yet, but may have some instances in the year 2030 (or never). As soon as an entity *m* satisfies the defining conditions of the class *mannedMarsMission*, it can be classified as *MannedMarsMission* and a new universal MannedMarsMission may emerge with the first instance of this class. Then, the class has instances that are in the extension of a universal:

$$MannedMarsMission \text{ EquivalentTo } ext(\text{MannedMarsMission}) \quad (5)$$

The class *MannedMarsMission* can be used in new definitions, e.g.

$$MannedMarsMissionPlan \text{ EquivalentTo } Plan \text{ and realizedBy only } mannedMarsMission \quad (6)$$

Here we define what a manned mars mission plan is, regardless of whether such a plan has ever been drafted, let alone realized. If desired, we could replace every defined term with its definiens, so that only classes which extend universals are explicitly named in the DL theory:

$$MannedMarsMissionPlan \text{ EquivalentTo } Plan \text{ and realizedBy only } (MarsMission \text{ and hasParticipant some } Human) \quad (7)$$

The latter definition would be in line with the principles of realist ontologies, because it does not name any non-referring unit.

These examples show that the application of DL operators to a set of classes that are the extensions of universals allows the straightforward generation of classes that do not represent universals. These classes can then be used for inferences. For example, from the def-

inition of *MannedMarsMissionPlan*, together with the definition:

MarsMissionPlan **EquivalentTo** *Plan*
and realizedBy only *MarsMission* (8)

we can infer that

MannedMarsMissionPlan **SubClassOf**
MarsMissionPlan (9)

While mars missions are scientifically feasible, unicorns are scientifically implausible. Nevertheless, a psychiatrist may need to use the term “unicorn” when describing the topic of delusional disorder of patients who feel themselves persecuted by unicorns.

In contrast to a realist ontology, a DL theory may contain a defined class *unicorn*:

unicorn **EquivalentTo** *Horse* **and hasPart**
some *Horn* (10)

without stating that it has instances. Such a statement asserts that, if there are ever horses with horns, then they would be classified as unicorns. This demonstrates that classes in a DL theory may specify the meaning of terms within a vocabulary without any commitment to whether the terms in the vocabulary refer to universals or not.

When describing a unicorn delusion we can refer to the class *unicorn* by a universal restriction without asserting any existence of instances of *unicorn*, e.g.

UnicornDelusion **EquivalentTo** *Delusion*
and isAbout only *unicorn* (11)

We can again replace all defined terms with their definiens to name only those classes in the definition that we believe to represent universals:

UnicornDelusion **EquivalentTo** *Delusion*
and isAbout only (*Horse* **and hasPart**
some *Horn*) (12)

This would then be compatible with the tenet of realist ontologies. We could even go further and define, within the realist framework, the class *PinkUnicornDelusion*:

PinkUnicornDelusion **EquivalentTo**
Delusion **and isAbout only** (*Horse* **and**
hasPart some *Horn* **and bearerOf**
some *PinkColor*) (13)

A description logics classifier then computes

PinkUnicornDelusion **SubClassOf**
UnicornDelusion (14)

because every instance of *PinkUnicorn* would be an instance of *Unicorn*, provided there were such instances.

It may not always be possible to avoid TDRs in ontologies. Therefore we may discuss whether a realist ontology implemented in DL should be allowed to have extensions for classes for which we do not yet know whether they have members. For example, whether the scientifically postulated Higgs bosons exist is one of the most exciting research questions in modern physics.

ScalarBoson **SubClassOf** *Boson* (15)

higgsBoson **SubClassOf** *ScalarBoson*
and *NeutralParticle* (16)

The latter axiom uses the intersection between the (non-empty) classes *ScalarBoson* and *NeutralParticle*. We can define the class

HiggsBosonResearchProject **EquivalentTo**
ResearchProject **and isAbout only**
higgsBoson (17)

which is then classified a subclass of, e.g. *BosonResearchProject*, defined as

BosonResearchProject **EquivalentTo**
ResearchProject **and isAbout only**
Boson (18)

These examples show how DL models can produce correct inferences using classes that are the extensions of TDRs. They also demonstrate how DL expressions with the same power can be built without such classes, thus corresponding to the principles of realist ontologies. An important modeling feature is the universal quantifier used with the relations **isAbout** and **realizedBy**. In opposition to the information artifact ontology (IAO) we do not claim that the ranges of these relations always have instances.

3.2 DL Theory Extended by a Taxonomy of Terminological Units

The above approach treats TDRs as classes in a DL theory which can be distinguished from classes extending universals within the formal theory itself. Although this may be sufficient for many use cases, it may be important in certain cases to explicitly state whether a class

is believed to represent a universal or not. Therefore, we will introduce a simple theory of terminological units which allows explicit references to TDRs. Such a theory requires an extension to the upper level and introduces a root class *TerminologicalUnit* that has terms as instances. These terms can then be explicitly asserted to have referents (in this world) or not, which makes it possible to distinguish between terms that represent universals and those that do not.

Our proposed extension uses a **hasInstance** relation that is rarely explicitly introduced in OWL theories. The main reason for this is that the notion of instantiation is commonly considered as part of DL semantics (in the form of class membership as determined by an interpretation function). However we here make a distinction between instantiation (in this world) and class membership. Whereas class membership relates a class to an individual, instantiation relates representational units (which are instances of the *TerminologicalUnit* class) to individuals.

Modifying the above formatting conventions, we use Underline for terms in a broader sense. Terminological units can refer to diverse flavors of universals, but also simple names or concepts (as entities of thought). The common characteristics of terminological units is that they have classes as their extensions. Classes are not required to have members. In contrast to sets, classes are defined intensionally. Therefore, empty classes are not necessarily identical. We introduce the following ground axiom: All members of the class *J*, which is the extension of a terminological unit *J*, are instances of *J*.

$$\forall x: \text{hasMember}(J, x) \Leftrightarrow \text{hasInstance}(\underline{J}, x) \quad (19)$$

Whereas the class instantiation is a built-in OWL feature, the explicit instantiation relation between a terminological unit and its instances needs to be manually asserted. Due to OWL restrictions this requires that terminological units are treated as individuals. We express this in DL as follows (with **instanceOf** being the inverse relation of **hasInstance**):

$$J \text{ EquivalentTo } \text{instanceOf value } \underline{J} \quad (20)$$

This is to be read: every member of the class *J* is an instance of some terminological unit *J*.

We therefore propose an extended DL theory with a bipartition into (i) *Particular* and (ii) *TerminologicalUnit*, with a taxonomy of terminological units under the latter. The terminological units that correspond to the classes below *Particular* are represented as individual members of the class *TerminologicalUnit*.

Let us illustrate this extended theory using the example *Animal*. AnimalTerm is an instance of the class *Universal*, which is a subsumed by *DenotingTerminologicalUnit* and *TerminologicalUnit*.

$$\underline{\text{AnimalTerm}} \text{ rdfs:type } \text{Universal} \quad (21)$$

$$\text{Animal} \text{ SubClassOf } \text{instanceOf value } \underline{\text{AnimalTerm}} \quad (22)$$

We can express classes that have no instances as follows:

$$\text{EmptyClass} \text{ EquivalentTo } \text{TerminologicalUnit} \text{ and not hasInstance some Thing} \quad (23)$$

Note that this is not equivalent to the class *Nothing*, which is a class that could not have any instances (via **rdfs:type**).

EmptyClass is therefore a class with an empty extension with respect to the explicitly introduced **hasInstance** relation. With this statement we can use an OWL reasoner such as Hermit [18] to ensure that this class is not instantiated (with respect to **hasInstance**), while it is at the same time possible to create subclasses and reason over them.

4 Implementation

The examples in section 3 were implemented in an OWL file with DL SROI expressivity (see <http://purl.org/steschu/misc/ICBO2011>).

Fig. 1 demonstrates the three terms PinkElephantTerm, PinkUnicornTerm, and UnicornTerm as members of the class *NonDenotingTerminologicalUnit*. The latter is defined as a terminological unit which has no instances. As a consequence of the axiom

$$(\text{instanceOf value } \underline{\text{UnicornTerm}} \text{ EquivalentTo } \text{Unicorn}) \quad (24)$$

the class *Unicorn*, by design placed under *Object*, becomes inconsistent (subsumed under *Nothing*).

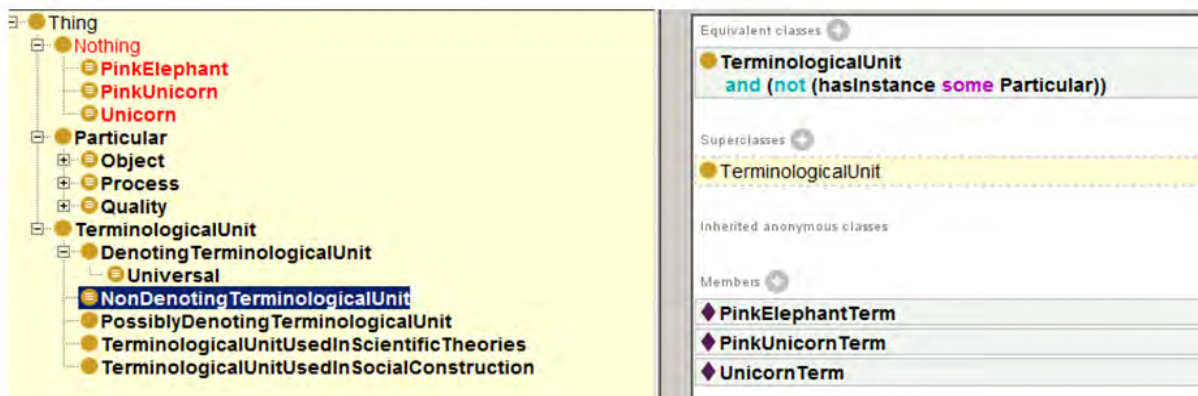


Figure 1. Example Ontology

5 Discussion

Fig. 1 illustrates our example ontology including a hierarchy of terminological units (Fig. 1). We consider the current arrangement of classes under *TerminologicalUnit* as preliminary and plan to harmonize it, in the future, with other proposals such as in [19]. So far, the following classificatory axes are proposed.

Possibility for existence. The first axis is the possibility for existence: is it logically possible for an instance of a subclass of *EmptyClass* to exist or not? Logical impossibility represents an unsatisfiable class, i.e., a logical contradiction. For example, all processes of raining and not-raining are logically impossible, because the definition of the class contains a contradiction. Arguably, the class of square circle is also inconsistent, given the background knowledge about “square” and “circle” is provided formally, and their mutual exclusiveness can be derived from the definitions and axioms. Unicorns, Higgs bosons or ideal humans, on the other hand, can possibly exist, only the world is such that they (probably) do not exist. In particular, they are not self-contradictory. Even unsatisfiable classes may be classified further. Square circles and rain-and-not-rain-events may both be unsatisfiable, yet they have different intensions [20]. It is, however, out of the scope of this paper to address differences in unsatisfiable classes.

Use in scientific theories. The second axis is its use in scientific theories. An entity can be predicted by a scientific theory which predicted other entities that we believe to exist. For example, Higgs bosons are predicted by a scientific theory that is compatible with a large

portion of collected empirical evidence. The problem is then to devise experiments and strategies to collect evidence for the existence of such predicted hypothetical entities.

Use in society, dependency on social construction. The third axis refers to idealizations as reference theories. These idealizations are often based on very similar entities that are observed, e.g., almost canonical humans.

6 Conclusions

Several biomedical ontologies have adopted the strategy to only include classes that either correspond to *universals* or are attributive collections, non-empty subsets of the extension of universals with an axiomatized inclusion criterion. The term “attributive collection” has been introduced recently to distinguish these classes in an ontological theory from the notion of “defined class” in a DL model [21]. In particular when ontologies are used for the representation of discourse, it becomes important to include classes that specify ideas about the terms used in scientific communication, independent of whether these terms refer to universals, attributive collections or not.

On the other hand, it may be relevant in some applications to include provenance information, including whether or not a class is believed to correspond to a universal or not. To provide this information, we started to develop a theory of terminological units and implemented this theory in OWL. Within this theory, terminological units, including both universals as well as classes that may not have instances in our world, are treated as instances of a

TerminologicalUnit class; and an explicit **hasInstance** relation is used to distinguish between universals (as terminological units that have instances) and empty classes (as terminological units that have no instances). The application of this theory will allow the development of ontologies that specify the meaning of terms within a vocabulary without sacrificing philosophical assumptions.

References

1. Michel Dumontier, Robert Hoehndorf (2010) Realism for scientific ontologies. Formal Ontology in Information Systems, Proceedings of the Sixth International Conference, FOIS 2010, 387-399.
2. National Center for Biomedical Ontology. Bioportal. bioportal.bioontology.org/
3. The Open Biological and Biomedical Ontologies. www.obofoundry.org/
4. Smith B et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-1255.
5. Basic Formal Ontology. ifomis.org/bfo.
6. DOLCE: a Descriptive Ontology for Linguistic and Cognitive Engineering, www.loa-cnr.it/DOLCE.html
7. BioTop. A top level ontology for the life sciences. purl.org/biotop/
8. GFO-Bio. A biomedical core ontology. onto.eva.mpg.de/gfo-bio.html
9. OBO Relation Ontology. obofoundry.org/ro/
10. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods of Information in Medicine, 37(4/5):394-403, 1998.
11. Schulz S, Johansson I. Continua in Biological Systems, 2007. The Monist 90 (4)
12. Smith, B. (2006). From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies. Journal of Biomedical Informatics, 39, 288-298.
13. Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. The Description Logic Handbook Theory, Implementation, and Applications (2nd Edition). Cambridge, 2007.
14. OWL2 Web Ontology Language. W3C. www.w3.org/TR/owl2-overview/
15. Ceusters W, Smith B. A unified framework for biomedical terminologies and ontologies. Stud Health Technol Inform. 2010;160(Pt 2):1050-1054.
16. Ruttenberg A. (2009) Defined classes. obi-ontology.org/page/Defined_classes
17. Horridge M, Patel-Schneider P (2009) OWL 2 Web Ontology Language Manchester Syntax. www.w3.org/TR/2009/NOTE-owl2-manchester-syntax-20091027/
18. Glimm B, Horrocks I, and Motik B. Optimized Description Logic Reasoning via Core Blocking. In Proc. of the 2010 Description Logic Workshop (DL 2010), volume 573 of CEUR, 2010. ceur-ws.org/
19. Ceusters W, Smith B. Foundations for a realist ontology of mental disease. J Biomed Semantics. 2010 Dec 9;1(1):10.
20. Loebe F, Herre H. Formal semantics and ontologies: Towards an ontological account of formal semantics. Formal Ontology in Information Systems, 49-62, 2008.
21. Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Applied Ontology 5 (2010) 139-188.

Towards an Ontology for Conceptual Modeling

James P. McCusker, Joanne Luciano, Deborah L. McGuinness

Tetherless World Constellation, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA

Abstract. Conceptual modeling can be viewed as a way of expressing human understanding of a body of knowledge. This view can be viewed as distinct from standard notions of data modeling and ontology, which seek to directly describe data and reality. We define our notion of conceptual interoperability, give use cases and requirements for it, and introduce the Conceptual Model Ontology (CMO), which satisfies the discussed use cases and requirements. We show how, using a common vocabulary, conceptual models can be used to tie together data at the level of conceptual interoperability. Finally, we introduce an implementation of CMO in the semantic web Biomedical Informatics Grid (swBIG), a linked data proxy for cancer Biomedical Informatics Grid (caBIG) models, semantic metadata, and data.

1 Introduction

The relationship between entities, the idea of the entities, and their information representation has come to the forefront of ontology and information modeling because conceptual models [1] have become critical in encoding human understanding of information [2]. To support this, layered representations of information models, such as the conceptual, logical, and physical model layers, [3] [4] have become common practice in many modeling disciplines. We seek a meta-modeling ontology that can easily express human understanding of entities and their data in terms of independent, reusable vocabularies that can be annotated onto conventional ontologies in a way that does not computationally disturb the annotated ontology (does not produce any undesired inferences) and does not require modification of the ontology or the data it represents. Our goal is to provide a way to use these sorts of annotations to satisfy certain use cases for conceptual interoperability [5].

2 Background

Biomedical ontologies have, for more than ten years, worked towards interoperability of data through use of verified categories, as has been the case in the Gene Ontology [6] and other OBO Foundry ontologies [7], reference models, as has been the case in Health Level 7 (HL7) Reference Information Model [8] and openEHR

[9], and common vocabularies, such as SNOMED-CT, LOINC, and NCI Thesaurus. However, integration across these ontologies has identified a number of challenges surrounding the strategies that were used to produce the ontology.

Ontologies are often created using one of two primary influences: linguistic or realist. Linguistic influences on ontologies stem from how people talk about and understand entities, whereas realist influences on ontologies stem from a focus on scoping by including only things that have scientific evidence of existence in the real world [10]. We refer to HL7-RIM as being linguistically influenced because it is primarily concerned with communication of human-generated records between entities. Ontologies with realist influence attempt to model reality as it is, and only model things for which there is scientific evidence [11]. The Basic Formal Ontology (BFO) [12] is possibly the most rigorous example of an ontology with realist influences. A realist strategy can provide a framework for relating other strategies, including conceptual models. Smith *et al.* [2] have developed a three-layer system for things in the world, our ideas of them, and representations of those ideas. Specifically, they define the following levels of entities that are involved in ontologies:

Level 1: the objects, processes, qualities, states, etc. in reality (for example on the side of the patient);

Level 2: cognitive representations of this reality (for example, on the part of researchers and others);

Level 3: concretizations of these cognitive representations in representational artifacts (for example, textual or graphical).

In a conceptual model, we consider Level 1 entities to be realist classes and properties. Level 2 entities can be classes and properties or concepts. Level 3 entities are terms, which are expressed as lexicographical labels for classes or concepts. We define a conceptual model as a set of Level 2 entities where each “represents” a Level 1 or Level 2 class or property [1]. Classes and properties that are also concepts can therefore represent themselves. We define logical models to be collections of classes and properties from either Level 1 or 2. Level 2 entities that are both classes or properties and concepts therefore exist both in the conceptual model (those assertions that treat them as concepts) and in the logical model (those assertions that treat them as classes or properties).

We seek to use conceptual models to achieve conceptual interoperability of data. Conceptual interoperability is the use of models of human understanding, or conceptual models, to provide interoperability commensurate with the level of alignment between conceptual models [13, 5, 14]. Two goals that we seek for conceptual interoperability are:

- Make similar but distinct data resources available for search, conversion, and inter-mapping in a way that mirrors human understanding of the data being searched.
- Make data resources that use cross-cutting models, such as HL7 v. 3 RIM¹ and provenance models (such as PML [15]), interoperable with domain-specific models without explicit mappings between them.

Resources such as the Gene Expression Omnibus (GEO) [16], ArrayExpress [17], and caArray [18] all contain separate logical models,

but rely on related conceptual models, MAGE for ArrayExpress and caArray [19] and MIAME for GEO [20]. By encoding this model over each resource with a common vocabulary, it could then become possible to search across all resources using a single query, or easily convert data from one resource to another. Similarly, conceptual interoperability could enable the ability to search for patient history across domain-specific databases using queries that only talk about patient history, as we show in our Translational Research Provenance Vision [21] for biomedical experiments.

2.1 Relevant Ontologies and Frameworks

We leverage properties and classes from the BFO [10] and Information Artifact Ontology (IAO), [22] which are implementations of the scientific realist perspective on developing ontologies. We also leverage SKOS [23] as a basis for simple common vocabularies and associating conceptual models with those vocabularies. This work was based on practical issues surrounding mapping semantics from the cancer Biomedical Informatics Grid (caBIG) [24] into the semantic web. We have in the past worked on converting caBIG’s layered semantics into OWL [25, 26] with success; however, the representation is limited to caBIG applications. Additionally, the mapping could not produce a one-to-one mapping between UML attributes and OWL properties, resulting in complex, unintuitive models.

3 Conceptual Interoperability Use Cases and Requirements

We divide the possible use cases of conceptual interoperability into three groups: search (or query), conversion, and direct mapping. Each of these use cases can be tailored to specific applications and additional requirements based on the level of interoperability needed. These use cases are necessarily abstract, and represent decompositions of uses cases such as testable hypothesis generation into component tasks.

Search: A user would like to perform queries with no knowledge of the underlying model. For example, “List the Education Level of all Persons in a dataset.” or “Find me all Tissue Specimens from Persons with an Adverse

¹ Health Level 7 Version 3 Reference Information Model [8]

Event while taking Drug *x*.” That “Drug *x*” is actually a class of drugs should not have to be a concern to the user.

Conversion: A user would like to convert instance² data from one logical model to another with a certain level of fidelity. This can be between domain models, or between a domain model and a cross-cutting model, such as a provenance model. For example, when events of *Clinical Service* occur with a given *Date*, dynamically create a record of *Vital Status* of *Alive* on that *Date*. These data are critical for tools like Kaplan-Meier survival curves [27], but availability of encounter data can be scattered across multiple organizations and systems that use different internal models.

Mapping: A user would like to create an automated mapping between two logical models. For example, take existing caBIG data models and align them with the BRIDG (Biomedical Research Integrated Domain Group) model [28]. This would occur when it is desirable for the Annotation and Image Markup [29] class *Person* to be automatically mapped as subclass of *bridg:Person*³ because of their mutual relationship with *ncit:Person*⁴.

We have identified a number of requirements for tools that would support these use cases:

Common Vocabulary: Conceptual models must use a common vocabulary that is distinct from any particular conceptual model. This is to allow portability of vocabularies between models, and prevent the reliance on one particular representation that might favor one logical model over another.

Distinction from Logical Models: A conceptual model and its vocabulary must not be represented in the same metamodel as a logical model. Doing so in metamodels that support reasoning may allow for direct inferences between conceptual and logical layers. This can have unintended consequences, for example, in cases where the logical and

conceptual models are both expressed in OWL. If the logical model has classes that are subsumed by conceptual model classes, then it no longer becomes clear whether the instance is referring to an instance of a thing, or an instance of an idea of a thing.

Natural, Idiomatic Expression: A conceptual modeling framework must support natural, idiomatic expression of the actual data in its natural form. This means that there must never be any need to modify a logical model or its data in order to allow annotation of a conceptual model onto it.

Types, Properties, and Relations: A conceptual modeling framework must provide a way to express relationships between types, properties, and relations.

Additional Relationships: Most concepts have inter-relationships that can assist in improving conceptual interoperability. Any framework must provide a way of expressing these additional relationships.

These requirements come from previous experience with modeling layered semantics using OWL [25] where relating models with reference terminologies expressed in the same language (OWL 1) proved problematic.

4 The Conceptual Model Ontology

The Conceptual Model Ontology (CMO)⁵ is a metamodel for representing conceptual models and their inter-relationships to logical models and vocabularies. Core to the CMO are these three classes:

cmo:Type. An abstract or general idea inferred or derived from specific instances, representing a set of those instances.

cmo:Quality. The conceptual representation of anything that is a property (a thing that is inherent in an entity, like eye color) or an attribute (a thing that has been assigned, or attributed, to an entity, like name or identification number). *cmo:Quality* is the union of those two sets, so issues relating to determining if a quality is an attribute or property are not relevant here.

² Instances here and in the rest of the paper informally refer to OWL Individuals, in particular, Type 1 individuals in reality.

³ BRIDG: Biomedical Research Integrated Domain Group. <http://bridgmodel.org>

⁴ ncit: NCI Thesaurus. <http://ncit.nci.nih.gov>

⁵ <http://purl.org/twc/ontologies/cmo.owl>

cmo:Relation. A concept representing the relationship between two independent entities.

Each of these classes are subclasses of *skos:Concept*, which is in turn asserted in CMO to be a subclass of *iao:information content entity* [22]. These concepts are considered Level 2 entities from Smith *et al.* [2]. Concepts are tied to logical model entities through the *cmo:represents* property, a subproperty of *iao:is about*. Entities in logical models can either be concepts or Universals (Type 1 entities). *cmo:Universal* is a subclass of *bfo:independent_continuent* and *cmo:Fiat Entity* is a subclass of *bfo:generically_dependent_continuent*. Both *cmo:Universal* and *cmo:FiatEntity* have requisite classes, qualities, and relations, and are intended to be types that are punned on to OWL classes and properties in the OWL 2 metamodeling pattern [30]. The class hierarchy is displayed in Figure 1. Classes that are not universals are usually considered to be themselves concepts, and are metamodelled as *skos:Concepts*. These classes, since they are themselves concepts, in a very real sense represent themselves. However, it is impossible for a universal to represent itself, since universals are not considered to be concepts.

Subproperties of *skos:broadMatch* are provided to provide relationships between CMO concepts and common vocabularies. We provide *cmo:hasPrimaryConcept* and *cmo:hasQualifier* to allow for more nuanced composition, for example allowing “Tissue Specimen” to have a primary concept of “Specimen” and a qualifier of “Tissue”.

We use SKOS as a basis for CMO because of its following properties. SKOS concepts unambiguously align to the definitions of concepts that we are using (as Level 2 entities), while OWL is ambiguous in its definitions of “class”, it could either be considered a set or a concept. This is important, because we seek to draw a distinction between concepts as they exist in conceptual models, and the sets of things that they represent. Alternatively, remaining in OWL DL means that to use OWL classes as a common vocabulary would mean either creating instances of that class or punning that class to an instance. Punning the class means that the instance no longer has any semantics associated with it, and would

need to either be given the type of the OWL class to regain semantics, or be given secondary semantics using an alternative structure. Here we do exactly that by giving the instance semantics using SKOS. Giving the class as a type of the instance in the conceptual model is also problematic, because it conflates being a thing of a type and being the idea of a type. The idea of a cat is not a cat, and when creating a conceptual metamodel that integrates with instance data, it is important to maintain that distinction.

cmo:Type relates to *cmo:Quality* through the use of *cmo:hasQuality* and its inverse, *cmo:qualityOf*. Qualities can have *cmo:values CanBe* assertions which provide the set of possible values for that quality. *cmo:Relation* has source (*cmo:hasSourceRole*) and target (*cmo:hasTargetRole*) types which help describe how those entities are related. Taken together, these qualities and relations form the structure of a conceptual model. The relations of CMO are outlined in Figure 2. By tying into existing common vocabularies, CMO-based concept models can be easily aligned along those vocabularies, as we will show below.

5 Implementation

The Conceptual Model Ontology is currently used as a backbone for “semantic web for the Biomedical Informatics Grid” (swBIG). This tool is currently available as a prototype RESTful service [31] that converts requests for resources from linked data URIs to caGrid service calls to requisite grid endpoints. This service uses a representation of NCI Thesaurus [32] converted to a SKOS representation using OWLtoSKOS. This representation addresses some, but not all of the concerns of Shulz *et al.* [33], and provides the ability to reason over concepts as instances in property value sets as well as in conceptual models. The retrieval operations are very simple and are documented on the swBIG web site. The source UML models are very closely mapped to preserve generalizations, attributes, and associations. Class and value typing on attributes and associations (using domain and range) and cardinality are preserved. When permissible values are listed for an attribute as part of the Common Data Element (CDE) [35] [36], an OWL ObjectProperty is created with a range of an enumeration class of the permitted

concepts (not strings). The concepts for classes, attributes, and properties as represented in CDEs are modeled using CMO. Instance data is generated using the model to determine and

query associations and convert values to concepts when a permissible value mapping is used.

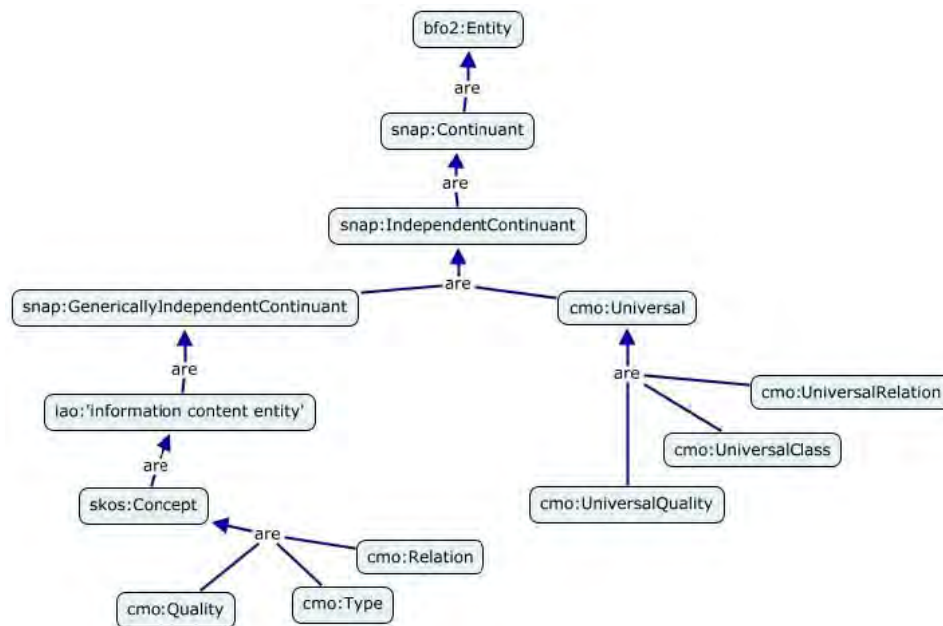


Figure 1. CMO Classes and their relationships with BFO, IAO, and SKOS. All diagrams are generated using CMap COE (<http://coe.ihmc.us>), and follow its labeling conventions.

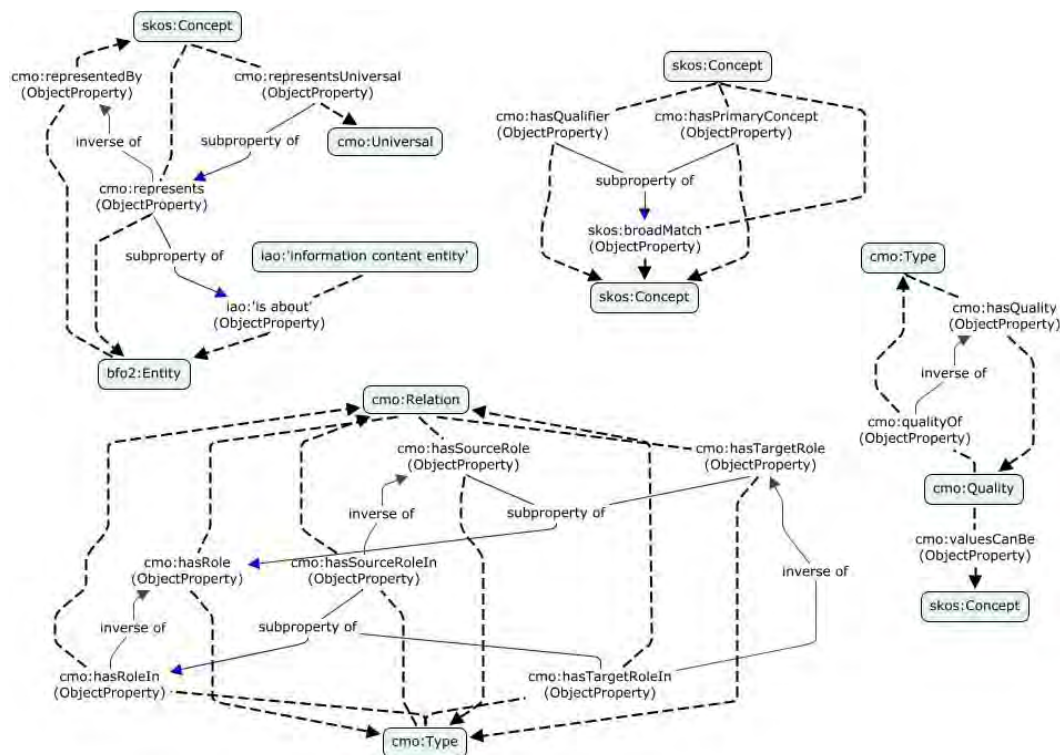


Figure 2. CMO properties and how they integrate with SKOS and IAO.

```

PREFIX cmo: <http://purl.org/twc/ontologies/cmo.owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#> PREFIX ncit:
<http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>

select count(distinct ?person) as ?count ?value
where {
  ?person a [cmo:representedBy [skos:broader ncit:Person]].
  ?person ?prop [rdfs:label ?value].
  ?prop cmo:representedBy [skos:broader ncit:Education_Level].
}

```

Figure3. Query 1: “Return the distribution of education level of all persons in the HINTS 2005 grid service.”
This query is performed by only using terms from a common vocabulary, NCI Thesaurus.

6 Evaluation

The Conceptual Model Ontology addresses all of the conceptual interoperability use cases and requirements. For example, the Query use case is satisfied with queries such as the one in Figure 3, which queries for the number of survey participants with a given level of education. The results of this query are displayed in Figure 4.

Conversion of data can be handled using rules such as the one in Figure 5. This example illustrates how CMO can be modified depending on the requirements of the task. The built-in semantics of CMO are kept minimal so that rules based on it can be tailored to the needs of the task. Some applications may require very strict conceptual alignment, while other applications may require a looser coupling in order to meet requirements. Models can also be mapped directly onto each other as shown in Figure 6.

CMO also satisfies conceptual interoperability requirements. Common vocabularies are distinct from the conceptual and logical models. Existing ontologies in OWL can be annotated without modification or change to existing semantics. While CMO is used to express semantics from caBIG, it is not limited to caBIG models. CMO provides a simple way to express relationships between types, properties, and

relations. Finally, because it uses SKOS-based common vocabularies, CMO allows additional relationships to be asserted between those concepts. For example, it is possible to assert that birds can fly at the conceptual level with direct assertions that have no automatic inference. Performing this using concepts means that this can be compared against instances without triggering consistency exceptions, such as the case with flightless or injured birds.

7 Future Work

We are currently investigating the use of CMO models to provide automated mappings of caBIG data elements into the BRIDG clinical model. This effort has seen some initial success, and work continues. Additionally, we will explore the use of CMO to represent domain-specific models in relation to a common model of provenance as envisioned in McCusker and McGuinness [21] including conceptual representations of biomedical experiments. We also are exploring the use of a common vocabulary to provide a unified view of existing provenance models and domain models in terms of provenance. We hope to do this with the Translational Medicine Ontology [37]. We plan to provide satisfaction of additional use cases as well.

Distribution of Educational Level in HINTS 2005 Survey

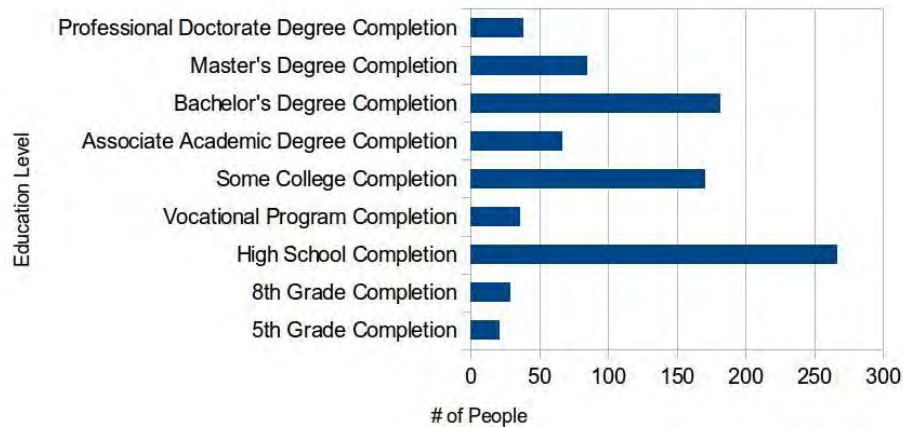


Figure 4. Distribution of educational level in the HINTS 2005 survey. These data were gathered using the query in Figure 3.

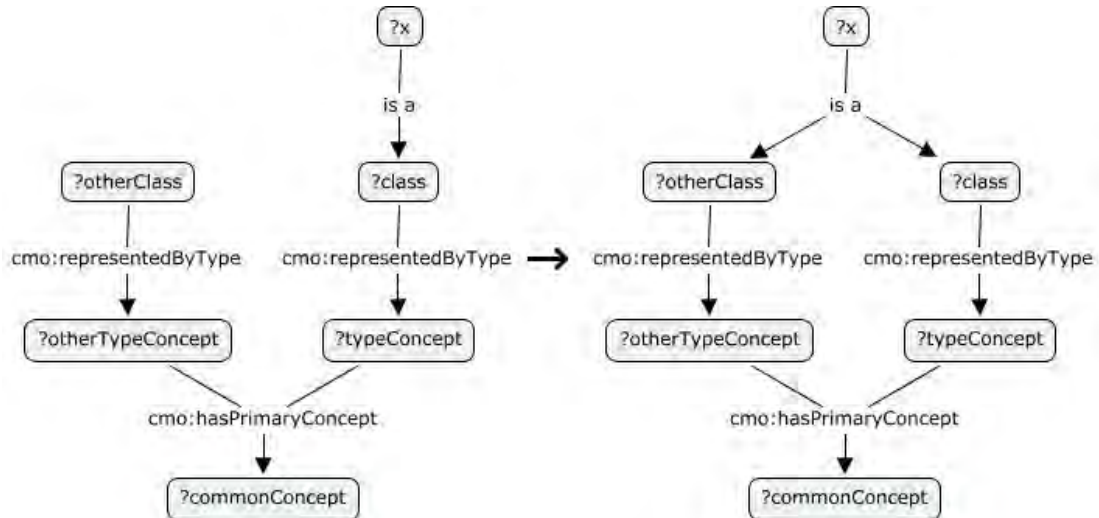


Figure 5. Mapping data from one logical model to another. By identifying that a “parent” class hangs directly off of a broader term of a “child” class, an instance of the “child” class can be given the type of the “parent”.

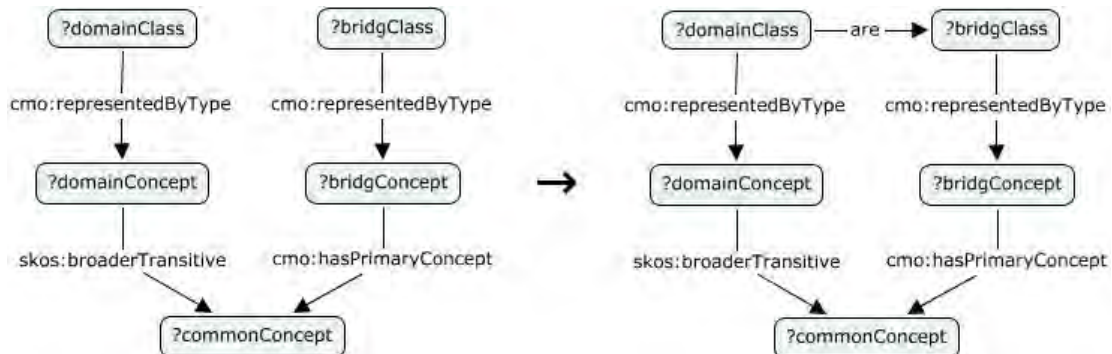


Figure 6. Mapping logical models directly on to each other can be accomplished by discovering relative relationships of the classes within the common vocabulary. The left hand side of the figure shows the precondition for mapping one class to another. The right side is the final state, where the added “are” arc represents the assertion that *?domainClass* is now a subclass of *?bridgClass*.

Finally, CMO does not yet provide a way to map between different levels of granularity. One model may represent a relationship as a direct link, while another may provide an intervening class which provides more information. It would be useful for CMO to include a property for how these levels relate.

8 Conclusion

Conceptual models can play a significant role in automated semantic interoperability, because they can allow the integration of data from across logical models without the need for direct integration of logical models. The Conceptual Model Ontology can support important uses cases in conceptual interoperability and is being used to represent existing semantics from a large software development program (caBIG). CMO is currently available for use with instance data using the swBIG linked data proxy. Finally, CMO is not limited to caBIG models, but can be applied to any logical model expressed in OWL.

References

1. Ceusters, W., Smith, B.: Foundations for a realist ontology of mental disease. *Journal of Biomedical Semantics* **1**(1) (2010) 10
2. Smith, B., Kusnierczyk, W., Schober, D., Ceusters, W.: Towards a reference terminology for ontology research and development in the biomedical domain. In: *Proceedings of KR-MED. Volume 2006.*, Citeseer (2006) 57–65
3. Melnik, S., Decker, S.: A layered approach to information modeling and interoperability on the web. In: *Proc. of the ECDL'00 Workshop on the Semantic Web.* (2000)
4. Luciano, J., Stevens, R.: e-Science and biological pathway semantics. *BMC bioinformatics* **8** (Suppl 3) (2007) S3
5. Tolk, A.: What comes after the semantic web-pads implications for the dynamic web. In: *Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation*, IEEE Computer Society (2006) 55
6. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al.: Gene ontology: tool for the unification of biology. *Nature genetics* **25**(1) (2000) 25
7. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology* **25**(11) (2007) 1251–1255
8. Schadow, G., Russler, D., Mead, C., McDonald, C.: Integrating medical information and knowledge in the HL7 RIM. In: *Proceedings of the AMIA Symposium*, American Medical Informatics Association (2000) 764
9. Kalra, D., Beale, T., Heard, S.: The openehr foundation. *Studies in Health Technology and Informatics* **115** (2005) 153–173
10. Smith, B.: Beyond concepts: ontology as reality representation. In: *Formal Ontology In Information Systems: Proceedings of the Third International Conference (FOIS-2004)*, IOS Press (2004) 73–84
11. Smith, B., Ceusters, W.: Ontological realism as a methodology for coordinated evolution of scientific ontologies. *Applied Ontology* **5** (2010) 139–188
12. Grenon, P., Smith, B.: SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition & Computation* **4**(1) (2004) 69–104
13. Tolk, A., Muguira, J.: The levels of conceptual interoperability model. In: *Proceedings of the 2003 Fall Simulation Interoperability Workshop*, Citeseer (2003) 007
14. Dobrev, P., Kalaydjiev, O., Angelova, G.: From Conceptual Structures to Semantic Interoperability of Content. *Conceptual Structures: Knowledge Architectures for Smart Applications* (2007) 192–205
15. McGuinness, D., Ding, L., Pinheiro da Silva, P., Chang, C.: Pml 2: A modular explanation interlingua. In: *Proceedings of AAAI. Volume 7.* (2007)
16. Edgar, R., Domrachev, M., Lash, A.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**(1) (2002) 207
17. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G., et al.: ArrayExpressa public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* **31**(1) (2003) 68
18. Bian, X., Klemm, J., Basu, A., Hadfield, J., Srinivasa, R., Parnell, T., Miller, S., Mason, W., Kokotov, D., Duncan, M., et al.: Data submission and curation for caArray, a standard based microarray data repository system. (2009)
19. Spellman, P., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., et al.: Design and implementation of microarray gene expression markup language (MAGE-ML).

- Genome biology 3(9) (2002)
20. Edgar, R., Barrett, T.: Ncbi geo standards and services for microarray data. *Nature biotechnology* **24**(12) (2006) 1471
 21. McCusker, J.P., McGuinness, D.L.: Explorations into the Provenance of High Throughput Biomedical Experiments. *Provenance and Annotation of Data and Processes* (2010) 120–128
 22. Ruttenberg, A., Smith, B., Ceusters, W.: *The Information Artifact Ontology* (2008)
 23. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. (2008)
 24. Von Eschenbach, A., Buetow, K.: Cancer Informatics Vision: caBIG. *Cancer informatics* **2** (2006) 22
 25. McCusker, J.P., Phillips, J., Beltrán, A., Finkelstein, A., Krauthammer, M.: Semantic web data warehousing for caGrid. *BMC bioinformatics* **10**(Suppl 10) (2009) S2
 26. Gonzalez-Beltran, A.: *Ontology-based Queries over Cancer Data*. (2010)
 27. Efron, B.: Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association* **83**(402) (1988) 414–425
 28. Fridsma, D., Evans, J., Hastak, S., Mead, C.: The BRIDG project: a technical report. *Journal of the American Medical Informatics Association* **15**(2) (2008) 130–137
 29. Rubin, D., Mongkolwat, P., Kleper, V., Supekar, K., Channin, D.: Annotation and image markup: Accessing and interoperating with the semantic content in medical imaging. *Intelligent Systems, IEEE* **24**(1) (2009) 57–65
 30. Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web* **6**(4) (2008) 309–322
 31. Richardson, L., Ruby, S.: *RESTful web services*. O'Reilly Media, Inc. (2007)
 32. Sioutos, N., Coronado, S., Haber, M., Hartel, F., Shaiu, W., Wright, L.: NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics* **40**(1) (2007) 30–43
 33. Schulz, S., Schober, D., Tudose, I., Stenzhorn, H.: The Pitfalls of Thesaurus Ontologization—the Case of the NCI Thesaurus. In: *AMIA Annual Symposium Proceedings. Volume 2010., American Medical Informatics Association* (2010) 727
 34. Hesse, B., Moser, R., Rutten, L., Kreps, G.: The health information national trends survey: research from the baseline. *Journal of Health Communication* **11** (2006) 7–16
 35. Warzel, D., Andonyadis, C., McCurry, B., Chilukuri, R., Ishmukhamedov, S., Covitz, P.: Common data element (CDE) management and deployment in clinical trials, *American Medical Informatics Association* (2003)
 36. Kunz, I., Lin, M., Frey, L.: Metadata mapping and reuse in caBIG. *BMC bioinformatics* **10**(Suppl 2) (2009) S4
 37. Dumontier, M., Andersson, B., Batchelor, C., Denney, C., Domarew, C.: The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench. In: *Proceedings of the 13th ISMB2010 SIG Meeting” Bio-Ontologies*. (2010) 120–123

What's in an 'is about' Link?

Chemical Diagrams and the Information Artifact Ontology

Janna Hastings^{1,2}, Colin Batchelor³, Fabian Neuhaus^{4,5}, Christoph Steinbeck¹

¹Chemoinformatics and Metabolism, European Bioinformatics Institute, Cambridge, UK

²Swiss Center for Affective Sciences, University of Geneva, Switzerland

³Informatics, Royal Society of Chemistry, Cambridge, UK

⁴National Institute of Standards and Technology, Gaithersburg, MD, USA

⁵University of Maryland Baltimore County, MD, USA

Abstract. The Information Artifact Ontology is an ontology in the domain of information entities. Core to the definition of what it is to be an information entity is the claim that an information entity must be 'about' something, which is encoded in an axiom expressing that all information entities are about some entity. This axiom comes into conflict with ontological realism, since many information entities seem to be about non-existing entities, such as hypothetical molecules. We discuss this problem in the context of diagrams of molecules, a kind of information entity pervasively used throughout computational chemistry. We then propose a solution that recognizes that information entities such as diagrams are expressions of diagrammatic languages. In so doing, we not only address the problem of classifying diagrams that seem to be about non-existing entities but also allow a more sophisticated categorisation of information entities.

Introduction

As the importance of ontology in biomedicine grows, the attention of ontologists is being pressed to the tasks of disambiguation of domain terminology and clarification of underlying hierarchies and relationships in an ever-wider network of interrelated domains [2, 10]. Some issues are emerging as similarly problematic in many of these different domains. One such is the clear definition and distinction of foundational types such as *processes* and *dispositions* [1]. Another is the confusion between *information entities*, such as computer simulations, models and diagrams, and the entities that they are models and diagrams of. It is to this latter problem that we turn in this paper.

Chemical graphs are the molecular models that are used throughout chemistry to succinctly describe chemical entities and allow for computational manipulations [12, 6]. Chemical graphs are typically depicted graphically as schematic illustrations – chemical diagrams. Chemical graphs and chemical diagrams are examples of information entities in the chemical domain, and their use has become so pervasive that language used by chemists to refer to chemicals regularly interchanges words for information (such as 'graph') with words for

actual chemicals [6].

The Information Artifact Ontology (IAO) [8] is an ontology being developed for the domain of information entities of relevance in biomedicine. The fundamental criterion by which information entities are defined and categorised in the IAO is their *aboutness*, that is, the types of entities that they are *about*. A diagram illustrating the chemical structure of caffeine molecules, for example, is about the class of caffeine molecules. While in this case the chemical diagram corresponds to something in reality (caffeine molecules), there are many other useful and scientifically relevant chemical diagrams that are not about something that exists. Thus, these chemical graphs are not information entities as currently defined in IAO. A similar scenario applies to many other models used in biomedicine, for example pathway diagrams and the mathematical models used in quantitative systems biology. Using chemical diagrams as examples, we will argue that information entities in IAO are defined too narrowly. Since information entities may not necessarily be about something, they cannot be categorized merely by what they are about. But, as we will argue, they should rather be categorised by what sort of information entities they are in their own right.

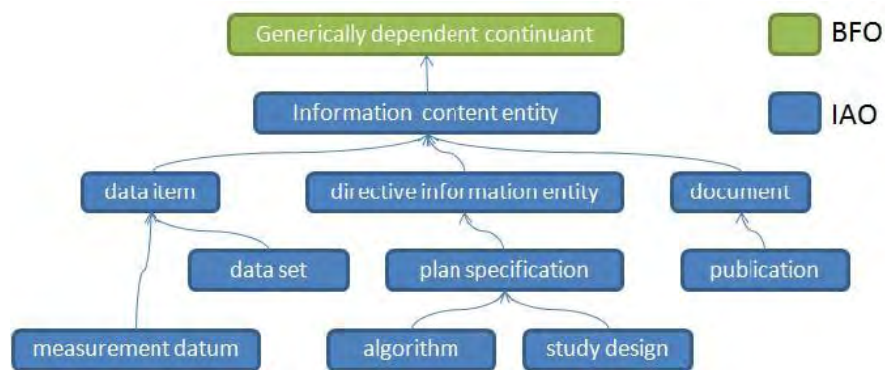


Figure 1. An overview of the Information Artifact Ontology

The remainder of this paper proceeds as follows. In the next section we briefly describe the IAO and the theory of chemical graphs and their related diagrams. Thereafter, we highlight the insufficiency of aboutness in defining types of diagrams. We go on to introduce some semantics for the representation relationship between chemical diagrams and chemical entities; and finally, we propose a modified approach to information ontology that is free of the problems with the current approach.

1 Background

1.1 The Information Artifact Ontology

The Information Artifact Ontology (IAO) [8] is an ontology of information entities being developed in the context of the Open Biological and Biomedical Ontologies (OBO) Foundry [9], beneath the upper level ontology Basic Formal Ontology (BFO) [11, 5]. Within this context, information entities are defined as:

Definition 1. *An information content entity (ICE) is an entity that is generically dependent on some artifact and stands in the relation of aboutness to some entity.*

The generic dependence on an artifact (i.e., a human creation) in the above definition restricts the scope of the domain to human-created information entities. The ‘generic’ part of the dependence captures the intuition that information can be copied, that is, reproduced in multiple bearers, in a way that hair colour, for example, cannot. The textual definition also refers to a relation of ‘aboutness’, which is further supplemented by the axiom:

$$ICE \text{ subClassOf } \text{is about some Entity} \quad (1)$$

The above is given in the Manchester Web Ontology Language (OWL) syntax, in which the existential quantification (\exists) is expressed using the infix some operator. This should not, however, obscure the strong existential dependency claimed, namely: for every *ICE*, there exists some entity to which the *ICE* is related by the **is about** relationship.

A hierarchical overview of the IAO together with some examples of information content entities (*ICEs*) is illustrated in Figure 1.

1.2 Chemical Graphs and Diagrams

The principal object of graph theory is a graph, which consists of a set of objects and the binary relations between them. Graph theory has found many applications in chemistry and is used to represent molecular entities through the molecular graph. These graphs represent the constitution of a molecule in terms of nodes (usually atoms, but in some cases groups of atoms) and edges (chemical bonds) [12].

For the purposes of this paper we define chemical graphs as follows¹.

Definition 2. *A chemical graph, denoted CG, is a tuple (V, E) in which each vertex $i \in V$ corresponds to an atom in a molecule; and each undirected edge $\{i, j\} \in E$ corresponds to a chemical bond between the atoms i and j .*

These CGs are based on the valence bond model of quantum mechanics [7]. For many of the molecules most relevant to the pharmaceutical industry this model reasonably accurately represents (1) by atoms, those

¹ We ignore additional complexity such as the representation of stereochemistry.

portions of the molecules that chemists associate with particular atoms, and (2) by bonds, those portions of the molecules that have high electron probability density. Cheminformatics software uses these to make useful predictions about the chemical properties of a molecule so represented and the physical properties of an ensemble of those molecules. They also enable the schematic representation of molecules in diagrams.

Definition 3. A chemical diagram, denoted *CD*, is a diagrammatic illustration of the information encoded in a *CG*, which follows an agreed diagrammatic syntax for the representation of the graph information.

Some examples of *CDs* are illustrated in Figure 2. In the 2D wireframe depiction, the diagrammatic syntax used specifies that the *CD* corresponds to the *CG* in that, for each edge $\{i, j\} \in E$ there is a corresponding line, and for each vertex $i \in V$ there is a corresponding *corner* or *line ending* in the *CD*. In the 3D ball and stick diagram, edges are illustrated with lines while vertices are illustrated with coloured, labelled spheres. In the 3D spacefill diagram, vertices are illustrated with large coloured spheres. Both the colours and the radii of the sphere are arbitrary – atoms are much too small to have colours, but the radii are based on experimental averages and are an approximation to the actual molecular structure.

Notice that there is not a one-to-one correspondence between *CDs* and *CGs*, since the same *CG* can be illustrated in many different *CDs*, obeying different syntaxes.



Figure 2. Some examples of *CDs* for the molecule caffeine

CDs, like maps, represent *spatial* information. Let us call spatial representations such as street maps, chemical diagrams, and engineering design models *structural diagrams* and, to a first approximation, assume that they have a direct structural association with a

portion of reality, which they are intended to represent.

Definition 4. A *structural diagram (SD)* is a diagrammatic representation of spatial aspects, such as position, topology and connection, of a structured portion of reality.

This definition, however, does not suffice, for reasons that will be described in the following section.

2 When ‘is about’ isn’t Enough

The agreed syntax of *CDs* allows their informational content to be reliably understood by all members of the community who use them for exchange of such information.

The agreed syntax also allows for the depiction of molecules, which are

1. Planned, in that the representation is used as a precursor to a synthesis procedure expected to produce a corresponding molecule instance.
2. Hypothesised, in that the representation corresponds to a molecule class for which it is not known whether corresponding instances exist.
3. Chemically infeasible, in that it is known that the representation illustrates a class of molecules for which no instances can exist for a measurable duration of time under normal conditions.
4. Impossible, in that the representation cannot be the structure of any molecule instances, since it violates the rules of molecular compositionality.

In the first two cases the *CD* might or might not be *about* molecules that exist. In the third case chemists expect, and in the fourth case they are certain, that the aboutness criterion of the IAO is violated. Nevertheless, these *CDs* are used by chemists to communicate and exchange information in the same ways as *CDs* that are known to correspond to something in reality. Thus, the way *CDs* are used does not justify treating only a subset of them as information entities. It also indicates that Definition 4 is not along the right lines.

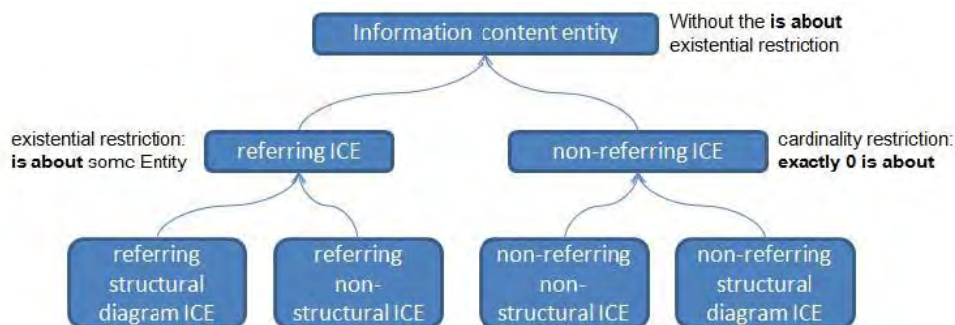


Figure 3. Referring and non-referring information entities in the IAO

A conceptualist resolution to this issue might defend a view of ontology as containing representations of *concepts*, and thereby not be required to differentiate between chemical diagrams for real or impossible molecules, or differentiate at the level of metadata only [4]. However, this seems to overlook the fundamental distinction between these cases, one that chemists recognise. Another strategy for addressing this problem is provided by Ceusters and Smith [3] who distinguish between *referring* and *non-referring* representational units in the context of a mental representation. The application of this distinction to an ontology of SDs beneath IAO is illustrated in Figure 3.

One obvious problem with this approach is that it leads to a massive level of parallel maintenance, since most types of *ICE* can appear twice in the ontology. A more fundamental objection is that this approach violates the fundamental design principles of BFO: categorization according to *ontological nature*, which does not change. For example, it is impossible for a tree (an independent continuant) to become a temporal region, or for a smile (a dependent continuant) to become a soccer game (an occurrent). However, according to the approach in [3] a *CD* might be a non-referring *ICE* now, but become a referring *ICE* tomorrow, because somewhere in some lab somebody accidentally synthesized the corresponding molecule. Thus, in contrast to the other ontological categories in BFO, it would be possible for non-referring *ICEs* to change their ontological nature. Even worse, the ontological nature of *CDs* would be affected by events that had no causal connection to the *CD* and did not change its structure in any way. Since the ontological nature of an entity is not affected by Cambridge changes, that is to say changes only in its description, we conclude that ‘non-

referring *ICE*’ and ‘referring *ICE*’ are not true ontological categories.

In summary, we agree with Ceusters and Smith that non-referring *ICEs* are *ICEs*. However, we reject the idea that the distinction between referring and non-referring should be the primary basis for classifying *ICEs*. There are some *ICEs* that are necessarily about something (e.g., photographs). But structural diagrams are information entities in virtue of the fact that they are well-formed *expressions* in a *diagrammatic language*. For each type of SD, there is a vocabulary (the symbols and icons that are used in diagrams of that type), a grammar that regulates how the elements of the vocabulary can be combined, and *compositionality* in the sense that the semantics of a complex expression is determined by the semantics of its components and the way these components are arranged.

The elements of the vocabulary of the diagrammatic language do need to correspond to something existing, otherwise the diagrams will not be scientifically relevant. However, not all combinations of the vocabulary that are permissible by the grammar will correspond to something in reality. It would seem strange indeed, on giving an ontological account of natural language, to divide all sentences into those that are about facts and those that are not. “Submariners love periscopes.” is a declarative sentence with a transitive verb regardless of whether it is a fact that submariners love periscopes. The same is true for expressions of diagrammatic languages.

3 The Ontology of Structural Diagrams

Different types of *CD* (such as 2D wireframe,

3D ball and stick) obey different diagrammatic syntaxes. What is essential to distinguish different types of diagrams is thus to provide a definition for these syntaxes.

Definition 5. A diagrammatic language $L_D = \langle V, G \rangle$ is an ordered pair that consists of the vocabulary V (a set of icons and symbols) and a syntax G of composition rules.

Definition 6. An interpreted diagrammatic language is a quadruple $IL_D = \langle V, G, T, \varphi \rangle$ such that $\langle V, G \rangle$ is a diagrammatic language, T is a set of types that is partitioned set of independent continuants IT and dependent continuants DT , and φ is a function that maps the elements from V onto T .

Definition 7. Let IL_D be an interpreted diagrammatic language as above, and let D be a well-formed expression in L_D (i.e., a diagram). D is a structural diagram that **is about** an entity x iff there is some injective interpretation function ι such that:

- for each element of V and each token t of V that is part of D , $\iota(t)$ is an instance of $\varphi(V)$
- for two tokens t_1, t_2 that are part of D and $\iota(t_1), \iota(t_2)$ are instances of elements of IT : t_1 is connected to t_2 iff $\iota(t_1)$ is connected to $\iota(t_2)$
- for all tokens t, t_1, \dots, t_n : if $\iota(t)$ is an instance of some element of DT and $t_1 \dots t_n$ are all connected to t , then $\iota(t)$ inheres in $\iota(t_1) \dots \iota(t_n)$.
- there is no part y of x such that y is an instance of some type in T and for all t that are part of D there is no $\iota(t) = y$.

Chemical diagrams of hypothetical molecules that do not exist are not about anything, but they are still well-formed expressions of an interpreted diagrammatic language. For example, the vocabulary V of the 3D ball and stick language consists of colored spheres and lines. The syntax G describes how these elements can be combined to diagrams. The set IT consists of types of atoms, the set DP consists of the types of chemical bonds that connect atoms within a molecule. The function φ maps the color-coded balls to types of atoms and the links to types of bonds. The second diagram in Figure 2 is a structural diagram of a given instance of a caffeine molecule x , since it is possible to map the spheres of the diagram to

the atoms that are part of x and the links of the diagram to the chemical bonds of x such that the connections in the diagrams corresponds to the chemical reality in the molecule. Conversely, if the diagram contains a link that does not correspond to a bond in a given molecule x or if it contains a sphere that is mapped to a type of atoms that do not occur as part of x , then the diagram does not represent x .²

To place SD s (and therefore CD s) as subtypes of IAO's ICE , we need to change the fundamental aboutness criterion from Equation (1) to a *value* rather than *existential* restriction:

$ICE \text{ subclassOf is about only Entity}$ (2)

This restriction no longer expresses an existential dependence. Rather, it now has the effect that *if* there is some entity that the ICE is about, *then* it must be of the required type to avoid a logical inconsistency. Note that this formula expresses a schema, which will be made more precise for different types of ICE . With the inclusion of **conforms to** axioms to relate the ICE to the L_D , we are now in a position to provide a better definition for SD s and CD s to replace Definition 4:

$SD \text{ subclassOf } ICE \text{ and is about only}$
 $StructuredEntity$
 and **conforms to** some
 $DiagrammaticLanguage$
 $CD \text{ subclassOf } SD \text{ and is about only}$
 $MolecularEntity$

We can safely include in the resulting ontology, illustrated in Figure 4, diagrams of planned, hypothetical, infeasible, and impossible molecules.

² The second clause of definition 7 is irrelevant in the case of CD s, because in CD s tokens of symbols for independent continuants (the atoms) are always connected by tokens of symbols for dependent continuants (the bonds). However, definition 7 is also intended to be applicable to diagrams where symbols for independent continuants might be connected directly; for example architectural drawings and engineering blueprints.

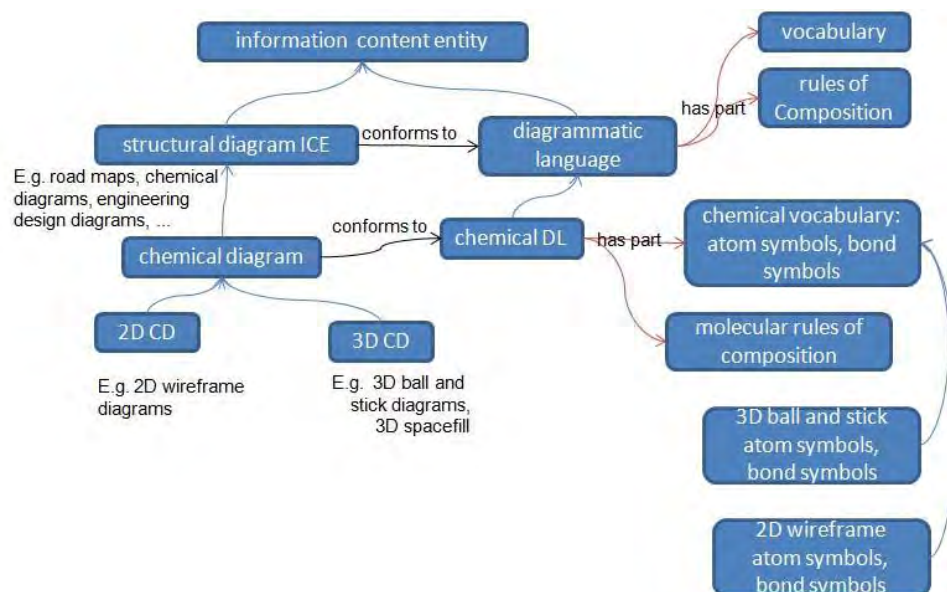


Figure 4. The ontology of chemical diagrams with distinctions for different syntaxes

Now, we can define different types of chemical diagrams regardless of their aboutness, and furthermore express the difference between different *types* of diagrams that are about the same entity (such as 2D and 3D diagrams of caffeine molecules). However, we can go one step further and define a *relationship* between 2D and 3D depictions of the same molecule.

Definition 8. Let L_1 , L_2 be two interpreted diagrammatic languages. Let θ_1 be a non-empty set of all well-formed expressions of L_1 , such that there is at least one diagram D in θ_1 and one entity x , such that D is about x in L_1 . Let θ_2 be a non-empty set of all well-formed expressions of L_2 , such that there is at least one diagram D in θ_2 and one entity x , such that D is about x in L_2 .

The function m is a coarsening from θ_1 (in L_1) to θ_2 (in L_2) iff

- m is a function from θ_1 onto θ_2 ; and
- for all diagrams D in θ_1 and all entities x : if D is about x in L_1 , then $m(D)$ is about x in L_2 ; and
- for all diagrams D_2 in θ_2 and all entities x :

if D_2 is about x in L_2 , then there is a diagram D such that D is about x and $m(D) = D_2$.

Coarsening functions map between two different diagrammatic languages, such that if a diagram in one language represents an entity, then it is possible to construct a diagram in the other language that also represents the entity. Typically, coarsening functions are *directed* from a greater to a lesser level of detail; that is, it is possible to map diagrams in a more detailed language to a diagram in a coarser language, but not the reverse. Coarsening functions allow us to define a relationship **coarser than** between SDs.

Definition 9. Let D_1 and D_2 be diagrams conforming to languages L_1 and L_2 , respectively. D_2 is coarser than D_1 iff

- there exists a function m and sets of diagrams θ_1 , θ_2 of L_1 and L_2 , respectively, such that m is a coarsening from θ_1 (in L_1) to θ_2 (in L_2) and $m(D_1) = D_2$; and
- there is no function m' such that m' is a coarsening from D_2 (in L_2) to D_1 (in L_1).

This is illustrated in Figure 5.

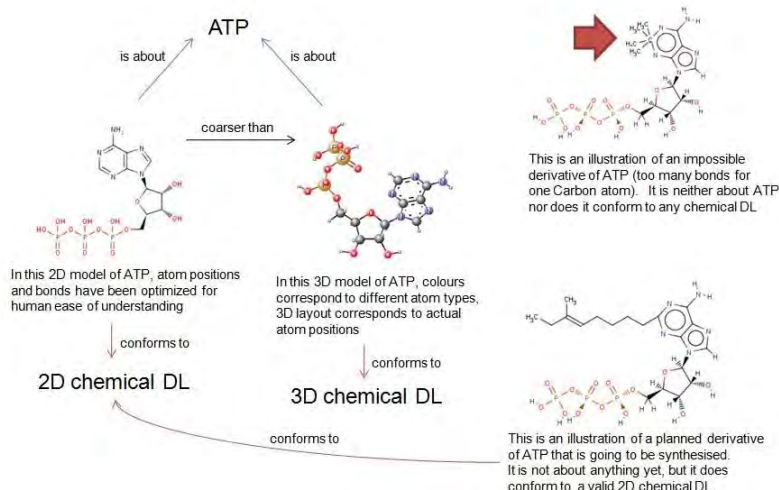


Figure 5. Some examples of chemical diagrams and their relationships

4 Conclusion

We have argued that the **is about** relationship is not enough to define *CDs*, for two reasons. Firstly, given the possibility of having several different *CDs* corresponding to the same molecule, we see that distinguishing between different types of diagrams, which obey different representational syntaxes, is not possible using only distinctions in what the diagram **is about**. Secondly, a challenge is posed in that *CDs* may be used validly to illustrate classes of molecules *for which no instances exist*. The existential dependency expressed in IAO means that the IAO cannot, in its present form, allow for the inclusion of such non-referring information entities.

We evaluated an approach based on parallel maintenance of IAO hierarchies with differing **is about** commitment. While such parallel maintenance may be a scientifically-valid strategy in some scenarios, it is unable to express the fact that the same representational formalism (i.e., diagrammatic syntax) is used across the hierarchies. Of course, the diagrammatic syntax, if it is to be scientifically-valid, must *typically* represent entities which do exist. But the syntax allows for compositionality and it would be absurd to require the existence of instances for all the complex expressions obtained by composing the elements of the representational vocabulary.

We therefore propose the definition of structural diagrams such as chemical diagrams based on their syntaxes. Any diagram

expressed in an interpreted diagrammatic syntax is a valid information content entity regardless of the existence of instances that the diagram **is about**; although the existence of such an instance may be an interesting property depending on the application scenario.

References

1. Batchelor, C., Hastings, J., Steinbeck, C.: Ontological dependence, dispositions and institutional reality in chemistry. In: Galton, A., Mizoguchi, R. (eds.) Proceedings of the 6th Formal Ontology in Information Systems conference. Toronto, Canada (2010)
2. Bodenreider, O., Stevens, R.: Bio-ontologies: current trends and future directions. Briefings in Bioinformatics 7(3), 256–274 (2006)
3. Ceusters, W., Smith, B.: Foundations for a realist ontology of mental disease. Journal of Biomedical Semantics 1(1), 10 (2010)
4. Dumontier, M., Hoehndorf, R.: Scientific realism. In: Galton, A., Mizoguchi, R. (eds.) Proceedings of the 6th Formal Ontology in Information Systems conference. Toronto, Canada (2010)
5. Grenon, P., Smith, B., Goldberg, L.: Biodynamic ontology: Applying BFO in the biomedical domain. In: Stud. Health Technol. Inform. pp. 20–38. IOS Press (2004)
6. Hastings, J., Batchelor, C., Steinbeck, C., Schulz, S.: What are chemical structures and their relations? In: Galton, A., Mizoguchi, R. (eds.) Proceedings of the 6th Formal Ontology in Information Systems conference. Toronto, Canada (2010)
7. Pauling, L.: The shared-electron chemical bond. Proc. Natl. Acad. Sci. USA 14, 359–362 (1928)

8. Ruttenberg, A., Courtot, M., The IAO Community: The Information Artifact Ontology (2010), <http://code.google.com/p/information-artifact-ontology/>
9. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25(11), 1251–1255 (Nov 2007)
10. Smith, B., Ceusters, W.: Ontological realism as a methodology for coordinated evolution of scientific ontologies. *Applied Ontology* 5, 139–188 (2010)
11. Smith, B., Grenon, P.: The cornucopia of formal ontological relations. *Dialectica* 58, 279–296 (2004)
12. Trinajstić, N.: Chemical graph theory. CRC Press, Florida, USA (1992)

Bioassay Ontology to Describe High-Throughput Screening Assays and their Results

Uma Vempati¹, Ubbo Visser², Saminda Abeyruwan², Kunie Sakurai¹, Magdalena Przydzial¹,
Caty Chung¹, Robin Smith¹, Amar Koleti¹, Christopher Mader¹, Vance Lemmon³, Stephan Schürer^{1,4}

¹Center for Computational Science, ²Department of Computer Science,
³The Miami Project to Cure Paralysis, Department of Neurological Surgery,
⁴Department of Molecular and Cellular Pharmacology,
University of Miami, Miami, FL, USA

Abstract. Huge amounts of high-throughput screening (HTS) data are generated in the pharmaceutical industry and more recently in the public sector. These are typically analyzed on a per-project basis. Comparison and analysis across many diverse HTS datasets are hindered by the lack of standardized descriptions of biological assays and screening results. Here, we present the BioAssay Ontology (BAO), which enables the categorization of biological assays by concepts relevant to interpret and compare HTS data and thus facilitates data analysis across many HTS campaigns. We used BAO to annotate assays from the largest public HTS data repository, PubChem. Here we demonstrate how BAO can be applied to access and analyze HTS data. BAO makes use of expressive description logic and has potential for discovering implicit knowledge using inference. BAO is publically available from the NCBO BioPortal at <http://bioportal.bioontology.org/ontologies/44531>.

Keywords: high-throughput screening, HTS, biological screening, assay ontology, bioassay ontology, description logic, bioassay, biological assay, semantic integration, data analysis

1 Introduction

High-throughput screening (HTS) has become the most commonly used approach to identify starting points for the development of novel drugs [1]. Increasingly complex biological systems and processes can be interrogated using HTS, leveraging innovative assay designs and new detection technologies. Driven by the NIH Molecular Libraries Initiative, HTS has become available to public research sector along with a public HTS data repository, PubChem [2]. Screening centers of the Molecular Libraries Probe Production Centers Network (MLPCN) have so far deposited thousands of HTS assays describing the effects of several hundred thousand compounds.

However, biological assays in PubChem currently lack standardized descriptions and standards to report the HTS results (endpoints). This hinders data analysis and integration, thus preventing researchers from utilizing these public resources to their fullest potential [3]. It is currently not possible to

identify related assays, for example those based on the same design (assay principle), the same detection technology, or interrogate protein targets from the same family or in the same pathway. It is also difficult to compare the activity of compounds across assays (because screening outcomes are not standardized). The motivation behind the BioAssay Ontology (BAO) development was to address this problem and to enable categorization of assays by concepts that are relevant to interpret screening results, which would then facilitate meaningful data retrieval and analysis across diverse HTS assays. We examined existing biomedical ontologies and incorporated information from several of them. However, we could not simply re-use or extend existing ontologies, because they lack many concepts required to model HTS data, including detection technologies (e.g. fluorescent vs. label-free assay), assay design (e.g. viability vs. enzyme reporter assay), HTS platforms, and detailed bioassay specifications. In addition, existing biomedical ontologies are not

structured to describe HTS assays by major categories, which describe important characteristics of an assay and whose subset of possible combinations could potentially define the universe of simple HTS assays. In an effort to develop a useful knowledge model of HTS assays, BAO uses description logic (DL) in OWL 2.0 to define semantics among classes of the different BAO categories. Although BAO represents an abstract model of HTS assays, its design and development adheres to many, although not yet all, principles recommended by the OBO (Biological and Biomedical Ontologies) Foundry [4, 5]. BAO can and should be mapped to additional ontologies of the OBO Foundry. For example the Ontology for Biomedical Investigations (OBI) [6] contains many classes that are related and relevant to describe biological assays. This work is currently in progress. We aim to collaborate with the members of the OBO to accomplish this while further developing BAO.

Here we describe the design and main components of BAO and the application of BAO for annotating HTS data in PubChem, both for data retrieval and meta-analysis. We also illustrate semantic definitions of BAO concepts and how they can be useful. BAO is publically available from our website [7] and the NCBO website [8].

2 Methods

We use ‘single quotes’ to denote terms from BAO and *italic* font to denote the semantic relationships.

2.1 Ontology Development

BAO was constructed using Protégé version 4.1 [9] in OWL (Web Ontology Language) 2.0 [10]. A number of available plugins were used throughout the development process including OWLViz2 [11] and OntoGraf [12] for visualization and DL reasoning engines HermiT [13] and Pellet [14].

2.2 PubChem Assay Annotation

To aid in manual annotation, assays were first clustered based on textual descriptions[15]. Using the terminology from BAO, PubChem bioassays were annotated with ~100 concept descriptors. These fall into the main BAO

categories ‘assay format’, ‘design’, ‘detection technology’, ‘meta target’, ‘perturbagen’, and ‘endpoint’. Assays were grouped by screening campaigns and organized by an assay stage (e.g. ‘primary’, ‘secondary’, etc.), throughput quality (e.g. ‘single concentration single measurement’, ‘concentration response multiple replicates’, etc.), and assay relationships (e.g. *is confirmatory assay of* or *is counter assay of*, etc.), among other categories. We also standardized assay endpoints, e.g.: ‘IC₅₀’, ‘EC₅₀’, ‘percent inhibition’, etc.

2.3 Integrating External Ontologies

We used OntoFox[16] to integrate external ontologies such as Gene Ontology (GO), NCBI Taxonomy, Cell Line Ontology (CLO) into BAO. Namespaces were preserved for these ontology terms.

3 Results

BAO is an abstract description of a ‘bioassay’ for the purpose of categorizing assays by concepts that are relevant to interpret screening results. BAO is therefore organized by major categories, which each include multiple levels of subclasses and specification classes. A number of specific object property relationships were created to connect classes and develop a knowledge representation in the domain of biological assays and screening outcomes. The relevance of BAO for annotating and analyzing MLPCN assays has recently been demonstrated [15]. Table 1 shows examples of attributes captured from BAO for sets of assays that comprise screening campaigns (partial views).

3.1 Ontology Design

BAO is instantiated in a well-specified syntax and designed to share a common space of identifiers. The ontology has a formally specified and clearly delineated content. All terms in the ontology have textual definitions.

Fig. 1 illustrates the high level design of BAO’s main components: ‘format’ and ‘perturbagen’ have direct relationships to ‘bioassay’; ‘meta target’, ‘detection technology’, ‘design’, and ‘endpoint’ are linked to ‘bioassay’ via ‘measure group’, which is an abstract concept to group experimental outcomes into

sets and thus to allow modeling multiplexed and multi-parametric assays. These classes are connected by specific (abstract) relationships (*has a* and its inverse *is of*) in contrast to the subsumption *is a* relationships within each major class.

One of the objectives of BAO is to represent domain knowledge, which is accomplished using DL in OWL 2.0. BAO (v1.1b868) has SROIQ(D) [17] expressivity and consists of 730 classes, 72 object properties (relations), 7 data properties, and 25 individuals (not including individuals from annotated assays or endpoints).

3.2 Description of BAO Major Concepts

The assay ‘format’ describes the biological or chemical features common to each test condition in the assay and includes several broad categories: ‘biochemical’, ‘cell-based’, ‘cell-free’, ‘tissue-based’, ‘organism-based’, and ‘physicochemical’ format. Further details are captured as ‘format specifications’ (Fig. 2) including special ‘reagent’ conditions (e.g.

‘redox reagent’, ‘detergent’, etc.) and ‘assay phase characteristic’ (‘homogeneous assay’ or ‘heterogeneous assay’).

Assay ‘design’ describes the assay methodology and implementation of how the perturbation of the biological system is translated into a detectable signal. In BAO, ‘design’ is broadly classified into one of eight categories: ‘binding reporter’, ‘enzyme reporter’, ‘inducible reporter’, ‘morphology reporter’, ‘viability reporter’, ‘redistribution reporter’, ‘conformation reporter’ and ‘membrane potential reporter’. All of these assay ‘design’ classes have further subclasses and specification classes (Fig. 2). BAO also describes the ‘detection technology’ used in bioassays, which includes ‘spectrophotometry’, ‘fluorescence’, ‘luminescence’, and others. Additional attributes of the assay technologies, such as standard screening kits (e.g. CellTiter-Glo) and pertinent parameters thereof are captured in the class ‘detection technology specification’.

Figure 1. BAO ontology excerpt showing root-level classes: ‘format’, ‘perturbagen’, ‘meta target’, ‘detection technology’, ‘design’, ‘endpoint’, and some of their relationships.

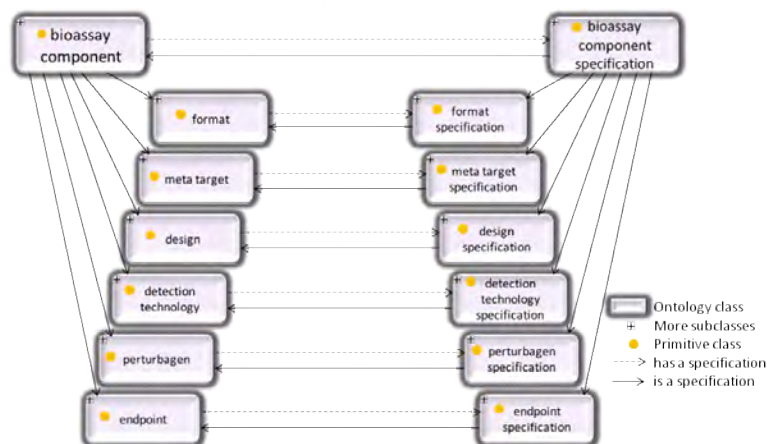
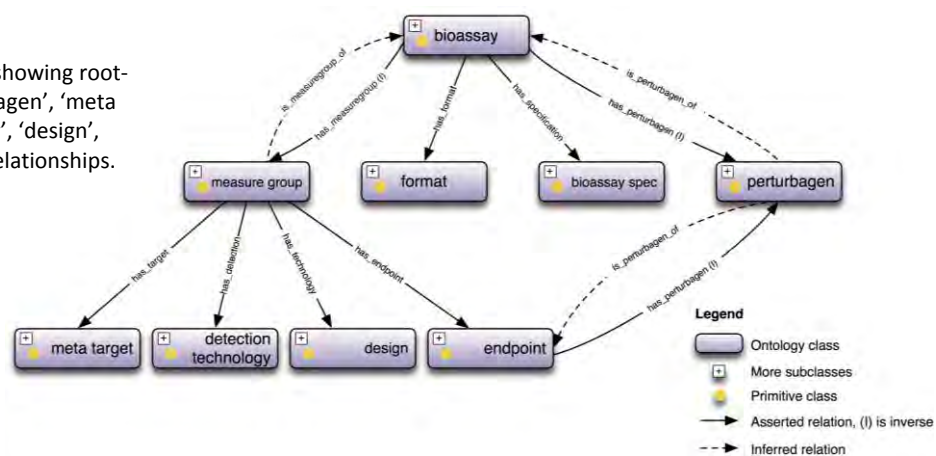


Figure 2. Illustration of the major bioassay components and corresponding specifications. Each bioassay component subclass has a corresponding specification subclass with a *has specification* and its inverse *is specification of* relationship.

Assay ‘meta target’ describes what is known about the biological system and / or its components interrogated in the assay (and influenced by the perturbagen). ‘Meta target’ can be directly described as a molecular entity (e.g. a purified protein or a protein complex), or indirectly by a biological process or event (e.g. phosphorylation). It includes information to enable linking external content, such as pathway databases, with the goal to infer the mechanism of action of perturbagens in an assay. The term “meta” is used to distinguish from the typical interpretation of target as a protein. Additional details about targets are captured as ‘meta target specification’. For example ‘protein specification’ (‘protein purity’, ‘protein form’, ‘protein preparation method’) or ‘cell specification’, which includes assay-specific details about the cell line (‘cell culturing component’, ‘cell modification’, ‘transfection specification’).

An assay ‘endpoint’ describes a quantitative or qualitative result of the bioassay. The main classes are ‘perturbagen concentration’ and ‘response’ endpoints, e.g. ‘IC₅₀’ or ‘percent inhibition’, respectively. Because ‘endpoint’ typically infers other information (e.g. mode of action: ‘inhibition’), in BAO the concept ‘endpoint’ is described semantically (using OWL DL) by specifying relationships between endpoints and other BAO concepts. The purpose is to enable the retrieval of inferred results that are not explicitly specified in a query (semantic equivalence) and which would otherwise not be retrievable or require complex Boolean endpoint queries (described in section 3.5).

3.3 Integrating BAO with External Ontologies

Some external ontologies contain information that define parts of concepts related to biological assays described by BAO (Fig. 3). We have imported relevant sections from Gene Ontology (GO) [18], Cell Line Ontology (CLO) [19], Unit Ontology (UO) [20] and others into BAO using OntoFox [16]. GO ‘biological process’ terms and CLO ‘cell line’ names and additional parameters are used in BAO ‘meta target’ and ‘meta target specifications’. CLO is currently being extended as a collaborative effort to cover cell lines relevant for biological screening [21]. Organism names associated

with targets were imported from NCBI taxonomy. Protein target names and IDs were referenced from UniProt. From UO we imported ‘concentration unit’ and ‘time unit’ terms. We also used terms necessary for curation of data from the Information Artifact Ontology (IAO), specifically, ‘information content entity’ and its sub-classes [22]. More work (in progress) is required to fully utilize IAO for BAO. We are currently working on mapping BAO to other OBO ontologies. For example, OBI includes relevant information to describe biological assays [6]. BAO was not developed as an extension of OBI because a different organization was required in BAO to allow categorization of assays and screening results for data retrieval and analysis. OBI links to many other resources and therefore mapping BAO to OBI will be of significant value as BAO and OBI take different but not incompatible approaches in describing assays. BAO can be seen as a more abstract description with the specific purpose to facilitate assay annotations by specific concepts and screening data analyses [15]; hence, many relationships are very specific (connecting two BAO classes; compare BAO design above). We have mapped some of the BAO relationships to the OBO Relationship Ontology (RO) and we aim to use more of RO relationships in the future. Additionally, we may be able to use RO to map BAO concepts to other ontologies.

3.4 Applications of BAO

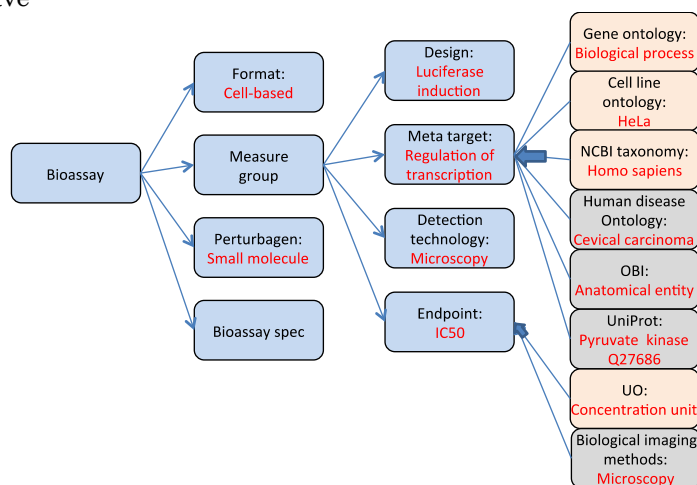
Using BAO terms, we have manually annotated over 900 MLPCN assays in PubChem, based on their textual descriptions (see methods). These assays correspond to over 15 millions endpoints (screening results), which we also standardized. Assays were categorized into campaigns. Examples of assay annotation are shown in table 1.

Using BAO annotations, assays can be readily categorized. For example, after annotating the formats of 1961 PubChem assays, we found the majority of assays to be ‘cell-based’ (874) or ‘biochemical’ (798) with some ‘organism-based’ (245) and a few ‘cell-free’ (31) and ‘tissue-based’ (13) (Fig. 4A). This information is relevant to interpret screening results, for example, ‘biochemical’ assays provide direct evidence of the mechanism of action (e.g. inhibition of an enzyme), while

activity in ‘cell-based’ assays infers that a compound is cell permeable. In another example, we annotated the most widely used HTS assay designs, namely, luciferase- and β -lactamase-based assays. 350 of such assays from PubChem were further classified into assay ‘design’ sub-categories (Fig. 4B). We have

demonstrated the utility of such annotations to identify promiscuously active compounds (i.e. their activity is related to the assay ‘design’, but not to the ‘meta target’) as well as the likely mechanism of action underlying such promiscuity [15].

Figure 3. Examples of external ontologies that contribute to some BAO concepts. External ontologies are shown to the far right and are linked to BAO concepts shown in blue to their left. The ontologies from which terms were already imported are shown in red and those that will be imported in the future are shown in grey. Specific examples of terms in an ontology or BAO concept are shown in red letters.



Screen. campaign	"Identification of inhibitors of Kruppel-like factor 5"		
AID	1700	1973	1944
Assay Stage	Primary Assay	Confirmatory Assay	Secondary Assay
Relationship	has confirmatory assay 1973 has secondary assay 1944	has primary assay 1700	has primary assay 1700
Assay Measurement	Single concentration single measurement (e.g. 1x%Inh)	Concentration response multiple replicates (e.g. 3xIC50)	Single concentration multiple replicates (e.g. 3x%Inh)
Throughput Quality	% Inhibition	IC50	% Inhibition
Endpoint std	% Inhibition	IC50	% Inhibition
Assay Format	Cell-based	Cell-based	Biochemical
Assay Design	Luciferase induction	Luciferase induction	Enzyme reporter
Assay Target	Transcription factor	Transcription factor	Hydrolase
Detection Technology	Luminescence	Luminescence	Luminescence
Measured Entity	Luciferase concentration	Luciferase concentration	Luciferin concentration
Screen. campaign	"Positive allosteric modulators of the M5 muscarinic receptor"		
AID	2665	2194	2206
Assay Stage	Primary	Confirmatory	Secondary Assay
Relationship	has confirmatory assay 2194 has secondary assay 2206	has primary assay 2665	has primary assay 2665
Assay Measurement	Single concentration single measurement (e.g. 1x%Inh)	Concentration response multiple replicates (e.g. 3xIC50)	Concentration response multiple replicates (e.g. 3xIC50)
Throughput Quality	Maximal response	EC50	EC50
Endpoint std	Maximal response	EC50	EC50
Assay Format	Cell-based	Biochemical	Cell-based
Assay Design	Calcium redistribution	Radioligand binding	Calcium redistribution
Assay Target	Human M5	Human M5	rat M1
Detection Technology	Fluorescence intensity	Scintillation counting, filter assay	Fluorescence intensity
Measured Entity	Calcium flux	[3H]-NMS radioactivity	Calcium flux

Table 1. Annotations of PubChem assays using BAO

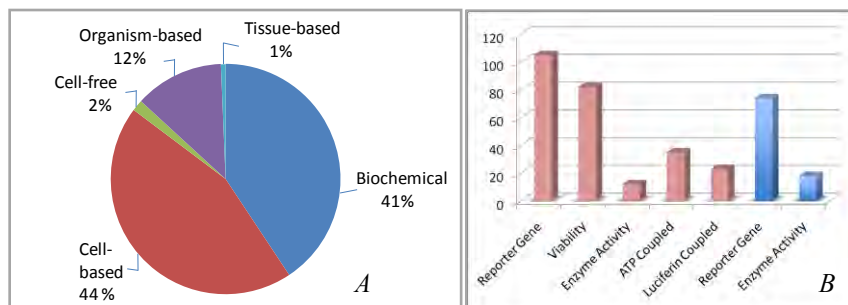


Figure 4.

Curation of PubChem bioassays.

A: Annotation of the bioassay ‘formats’ from PubChem.

B: Number of annotated PubChem assays by major assay platforms (luciferase (red) and β -lactamase (blue)) and relevant sub-categories.

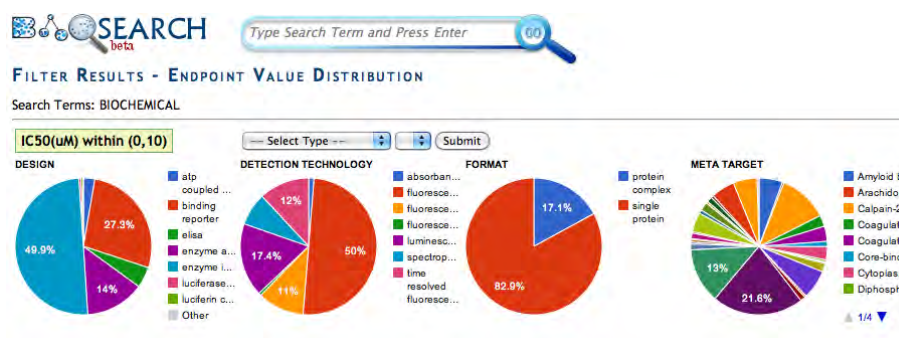


Figure 5. BAOsearch graphical summary example result page after querying for ‘biochemical’ assays (by concept search) with ‘IC₅₀’ endpoints of less than 10 micromolar activity. This page displays how the endpoints that match the search query are distributed among the major BAO concepts, ‘design’, ‘detection technology’, ‘format’ and ‘meta target’.

We have also developed a software application that makes use of BAO. BAOsearch [23, 24] allows us to query and explore annotated PubChem assays and screening results in the context of BAO (Fig. 5).

3.5 An Example of DL to Define BAO Concepts

To illustrate how BAO classes are embedded with semantic information, we describe the BAO ‘endpoint’ concept ‘IC₅₀’, defined as the concentration of the perturbagen that results in ‘50 percent inhibition’ (Fig. 6).

Definitions include equivalent classes (necessary and sufficient conditions) and superclasses (necessary conditions only). Necessary and sufficient conditions are used to classify individuals; for example, we might be able to infer that an individual endpoint must be an ‘IC₅₀’ because the ‘mode of action’ is ‘inhibition’ (among other criteria). With only necessary conditions, the definition is logically different, saying that if an individual is a member of the class ‘IC₅₀’, it is necessarily a subclass of ‘perturbagen concentration’. The equivalent class ‘IC₅₀’ specifies *has mode of action* only ‘inhibition’. “Only” denotes universal quantification, describing all the individuals whose *has mode of action* relationships refer to members of the class ‘inhibition’; or conversely, the individuals that do not have *has mode of action* relationships to individuals that are not members of the class ‘inhibition’. “Some” denotes existential restrictions, e.g. *has mode of action* some ‘inhibition’ specifies the existence of at least one relationship along a given property to an individual, which is a member of the class ‘IC₅₀’. Existential restrictions can be seen as “among other things”, and are used to “close” a given property, which is necessary for the reasoning

process (open world assumption). In the BAO knowledge model, many relationships specifically connect two classes and must not be interpreted in any other way. For example *has mode of action* means that ‘endpoint’ is further specified by ‘endpoint mode of action’. Certain specifications are inherited from classes that are higher up in the hierarchy. An example of this is the inherited anonymous class definition of individuals having the object property *has perturbagen concentration value*. There is also the relationship *has perturbagen*, describing that every individual of the ‘IC₅₀’ class must have at least one ‘perturbagen’. The semantic description of ‘IC₅₀’ facilitates the retrieval of inferred results. For example querying for perturbagens with greater than ‘50 percent inhibition’ at a defined ‘perturbagen concentration’ retrieves qualified ‘IC₅₀’ as well as ‘percent inhibition’ endpoints, as illustrated in the BAO SPARQL Examples on our web site [7]. (<http://129.171.150.121/joseki/query.html>)

Equivalent classes:

```
('has mode of action' some inhibition)
and ('has mode of action' only inhibition)
and ('has percent response' value '50
    percent inhibition individual')
```

Superclasses:

```
'has curvefit spec' only 'curvefit spec'
'perturbagen concentration'
```

Inherited anonymous classes:

```
('has perturbagen concentration unit' some
    'concentration unit')
and ('has perturbagen concentration unit'
    only 'concentration unit')
and ('has perturbagen concentration value'
    exactly 1 float)
('has specification' only endpoint spec)
('has perturbagen' some perturbagen)
and ('has perturbagen' exactly 1 Thing)
```

Figure 6. BAO definition of the class ‘IC₅₀’.

4 Summary

Large amounts of data are generated by HTS in private and public organizations. Nevertheless, large scale screening capabilities have so far not translated to increased numbers of approved drugs [25]. One likely reason is that available data is used inefficiently. It remains a challenge to effectively translate increasing amounts of data into actionable knowledge; at the very least this is the case for the current public domain data. To address this challenge, we have developed BioAssay Ontology (BAO). BAO describes biological assays and their outcomes by concepts that are relevant to interpret, analyze and integrate screening data. BAO addresses 1) development of standardized terminology and uniform standards to report HTS results; and 2) a semantic description of bioassays and their results to model domain knowledge and to facilitate semantic integration with diverse other resources [26, 27]. We have used BAO to annotate PubChem assays and showed that BAO concepts are useful to categorize and analyze screening results [24]. We have illustrated the use of DL to incorporate semantics into BAO concepts and to retrieve inferred query results.

BAO is under active development. Although BAO makes use of relevant information from several external ontologies, current effort is focused on incorporating content from several other resources with the potential of making BAO-annotated HTS data widely accessible via a "Linked Data" approach [28].

In summary BAO opens new functionality for querying and analyzing HTS data sets and has potential for discovering new knowledge (that is not explicitly described in the data) by inference.

Acknowledgements

The work presented here was supported by NIH grant RC2 HG005668. We acknowledge resources from the Center for Computational Science at the University of Miami.

References

1. Hertzberg, R.P., Pope, A.J.: High-throughput screening: new technology for the 21st century.

-
2. Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Bryant, S.H.: PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 37, W623-633 (2009)
3. Inglese, J., Shamu, C.E., Guy, R.K.: Reporting data from high-throughput screening of small-molecule libraries. *Nat Chem Biol* 3, 438-441 (2007)
4. Ashburner, M., Mungall, C.J., Lewis, S.E.: Ontologies for biologists: a community model for the annotation of genomic data. *Cold Spring Harb Symp Quant Biol* 68, 227-235 (2003)
5. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25, 1251-1255 (2007)
6. Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Soldatova, L.N., Stoeckert, C.J., Jr., Turner, J.A., Zheng, J.: Modeling biomedical experimental processes with OBI. *J Biomed Semantics* 1 Suppl 1, S7 (2010)
7. Center for Computational Science, University of Miami, <http://bioassayontology.org>
8. NCBO Ontology Portal, <http://bioportal.bioontology.org/ontologies/44531>
9. Noy, N.F., Crubezy, M., Fergerson, R.W., Knublauch, H., Tu, S.W., Vendetti, J., Musen, M.A.: Protege-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu Symp Proc* 953 (2003)
10. W3C, <http://www.w3.org/TR/owl2-overview/>
11. University of Manchester, <http://www.code.org/downloads/owlviz/OWLvizGuide.pdf>
12. Stanford University, <http://protegewiki.stanford.edu/wiki/OntoGraf>
13. Shearer, R., Motik, B., Horrocks, I.: HermiT: a highly-efficient OWL reasoner. In: 5th International Workshop on OWL: Experiences and Directions (OWLED 2008), pp. 10. Universität Karlsruhe, (2008)
14. Sirin, E., Parsia, B.: Pellet system description. In: Proceedings of the International Workshop on Description Logics (06). CEUR, (2006)
15. Schurer, S.C., Vempati, U., Smith, R., Southern, M., Lemmon, V.: BioAssay Ontology Annotations Facilitate Cross-Analysis of Diverse High-Throughput Screening Data Sets. *J Biomol*

- Screen 16, 415-426 (2011)
16. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. *BMC Res Notes* 3, 175 (2010)
 17. Horrocks, I., Kutz, O., Sattler, U.: The even more irresistible SROIQ. In: *Knowledge Representation (KR)*, pp. 57-67. AAAI, (2006)
 18. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet* 25, 25-29 (2000)
 19. Sarntivijai, S., Ade, A.S., Athey, B.D., States, D.J.: A bioinformatics analysis of the cell line nomenclature. *Bioinformatics* 24, 2760-2766 (2008)
 20. <http://biportal.bioontology.org/visualize/45500>
 21. Sarntivijai, S., Xiang, Z., Meehan, T., Diehl, A., Vempati, U., Schurer, S., Pang, C., Malone, J., Parkinson, H., Athey, B., He, Y.: Cell Line Ontology: Redesigning Cell Line Knowledgebase to Aid Integrative Translational Informatics. *ICBO conference paper* (2011)
 22. <http://biportal.bioontology.org/visualize/40642>
 23. Abeyruwan, S., Chung, C., Datar, N., Gayanilo, F., Koleti, A., Lemmon, V., Mader, C., Ogihara, M., Puram, D., Sakurai, K., Smith, R., Vempati, U., Venkatapuram, S., Visser, U., Schürer, S.: BAOSearch: A Semantic Web Application for Biological Screening and Drug Discovery Research. In: *Semantic Web Challenge, 9th International Semantic Web Conference (ISWC)*. (2010)
 24. Center for Computational Science, University of Miami, <http://baosearch.ccs.miami.edu/baosearch/>
 25. Mayr, L.M., Bojanic, D.: Novel trends in high-throughput screening. *Curr Opin Pharmacol* 9, 580-588 (2009)
 26. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G.T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the Semantic Web. *BMC Bioinformatics* 8 Suppl 3, S2 (2007)
 27. Smith, B., Brochhausen, M.: Putting biomedical ontologies to work. *Methods Inf Med* 49, 135-140 (2010)
 28. Bizer, C.: The emerging web of linked data. *IEEE Intelligent Systems* 87-92 (2009)

A Framework Ontology for Computer-Based Patient Record Systems

Chimezie Ogbuji

Case Western University (School of Medicine), Cleveland, OH, USA

Abstract. A lack of uniform content and format standards remains the biggest barrier to the development of the Institute of Medicine's (IOM) Computer-Based Patient Record (CPR). The CPR ontology is a uniform core set of data elements (whose formal semantics are captured in OWL) for use in a Computer-Based Patient Record (CPR). It is meant to be of immediate and practical value to clinical research projects and patient registries that wish to leverage well-defined ontologies for a given domain.

1 Introduction

In 1991, the IOM defined the Computer-Based Patient Record (CPR) as:

an electronic patient record that resides in a system specifically designed to support users by providing accessibility to complete and accurate data, alerts, reminders, clinical decision support systems, links to medical knowledge, and other aids [4].

This definition emphasizes data that are well-organized, discrete¹, and in a digital form.

One of the major goals of recent USA healthcare policy is to stimulate the *meaningful* use [2] of Healthcare Information Technology (HIT) in order to lay the groundwork for advanced electronic health information systems [7].

In describing the barriers to developing a CPR, the IOM declared that the content of CPRs must be defined such that each contains a uniform, core set of data elements that are named consistently. Some form of vocabulary control must be in place with specific meaning for data elements that describe clinical findings, clinical problems, procedures, and treatments [4]. The use of an ontology such as this one (the *CPR Ontology*) to govern a semantics for these data elements is meant to serve as an infrastructure skeleton to address this particular desiderata.

An ontology specifies a conceptualization of a domain and is often comprised of definitions of a hierarchy of concepts in the domain and restrictions on the relationships between them. An ontology with terms that serve as the basis for a uniform core set of data elements in a CPR will facilitate meaningful use of CPR content. In its current form, this ontology can be a high-level framework for collections of clinical vocabulary systems. All concepts are given a canonical syntax and semantics.

The scope of this ontology is over sub-domains of clinical medicine² as they relate to and appear in CPR content and with sufficient enough detail for immediate use in clinical care and research information systems. It can and has been used as a coordinate system for managing or eliciting precise meanings of terminology used by healthcare and research professionals for use in applications involving large scale data management.

To achieve maximal completeness of terminological coverage, this ontology leverages a number of normalization criteria that seek to minimize the amount of implicit differentiation between primitive concepts [11]. Definitions are given as chains of explicit restrictions on relations that differentiate two primitive concepts where one subsumes the other.

¹ The term discrete data is used to refer to data that are comprised of a finite set of values for variables. In the context of medical data, this is often referred to as coded data.

² The study of disease by direct examination of the living patient.

The CPR ontology	SNOMED-CT
clinical-finding	Clinical finding
procedure	Procedure
bodily-feature	Observable entity
organism	Organism
pharmacological-substance	Pharmaceutical / biologic product
recorded-clinical-situation	Situation with explicit context
clinical-artifact	Record artifact

Table 1. Mappings to SNOMED-CT

Like many contemporary ontologies that have a related goal, the CPR ontology is based on the Basic Formal Ontology. The following extant desiderata were followed in creating this framework:

1. Consistency with the top levels of the Basic Formal Ontology (BFO): reusing the terms *representational artifacts*, *occurents*, and *continuants* [17];
2. A clear separation of digital (“information content”) entities and the phenomenon they represent, or describe;
3. The use of a clinically-oriented framework for representing diseases, their causes, manifestations, related diagnostic acts, and pathophysiological phenomena. For these we leverage terms from the BFO, [13, 18], and SNOMED-CT.
4. A principled integration [12] with a reference ontology for (canonical) human anatomy.

2 Related Work

There are a number of issues with existing biomedical ontologies that the author has encountered while trying to harmonize this ontology with them. Contemporary ontology standardization and exchange is often hegemonic, but despite this, ontology harmonization and mapping can support interoperability. This ontology is meant to be used in this way with the contemporary BFO-based ontologies for clinical medicine and patient record content. In this section, we briefly summarize and characterize the overlap between this ontology and existing biomedical ontologies.

There are few principled ontology frame-

works or vocabularies that comprehensively meet the needs of a contemporary patient record or (disease) registry. In Bodenreider’s review [3] and analysis of high-impact biomedical ontologies, only one of the nine was listed with a scope of clinical medicine and patient record content: SNOMED-CT.

Neither the CPR ontology by itself nor many of the current BFO-based ontologies such as the Ontology of Biomedical Investigation (OBI), the Ontology for General Medical Science (OGMS), Biotop, or the Ontology of Information Artifacts (IAO) – by themselves (without extensions) – meet the IOM desiderata for use in patient record and registry systems in the manner that SNOMED-CT does. However, in sharing a common foundation, this ontology was designed to be semantically interoperable with them. Many existing biomedical ontologies don’t have a framework for representing the process of diagnostic inquiry and relationships with the information content that result from them. They also do not have a framework for the representation of a clinical diagnoses and the role of treatment in classifying diseases in a manner similar to that proposed in [19].

This ontology predates³ OBI (but not the MGED Ontology on which it is based), IAO, and OGMS. It has evolved as a result of interaction with the authors of many of them and is meant to facilitate harmonization for the benefit of users of both. A more comprehensive assessment of where these ontologies fail to provide coverage sufficient for use in detailed representation of the core data elements of a CPR is beyond the scope of this work.

The CPR ontology	The FMA
anatomical-structure	Material anatomical entity
immaterial-anatomical-continuant	Immaterial physical anatomical entity

Table 2. Mappings to the FMA

³ <http://www.w3.org/wiki/HCLS/POMROntology> (December 2006)

3 The Structure of the CPR Ontology

3.1 Running Example

Iuliano et al. [6] report the case of a young female patient with a previously clinically diagnosed Myocardial infarction (MI), a medical history of hypertension, no history of diabetes mellitus, hypercholesterolemia or premature coronary artery disease in her family. Ten years earlier, she had an inferior MI treated with systemic thrombolysis and other prescribed pharmacological substances.

During her hospitalization at that time, the echocardiogram revealed akinesis of the posterior-basal wall with an estimated ejection fraction of 50%. Laboratory tests including serum glucose and other requested components were performed. During the more recent hospitalization, a physical examination showed blood pressure of 135/80 mmHg, heart rate of 64 BPM, BMI of 27 kg/m², and waist circumference of 85 cm.

In subsequent sections, we will demonstrate how many of the clinical concepts involved in this case can be represented using the CPR ontology.



Figure 1. Diagram of clinical act taxonomy

3.2 Clinical acts

The **clinical-act** class corresponds to the root of a hierarchy of the actions that comprise the healthcare workflow. This hierarchy is placed directly under the *span:ProcessualEntity* class in the BFO ontology as shown in figure 1.

The recorded presence and absence of a

medical history is collected as a result of a **medical-history-screening-act**. These are a **screening-act** that results in anamnesis⁴. Other clinical acts that correspond to concepts in the case report are: **diagnostic-procedure** (echocardiogram), **laboratory-test**, **therapeutic-procedure** (thrombolysis), and **substance-administration** (prescription of the various drug regiments).

A **clinical-investigation-act** is distinguished from a clinical act by the restriction that they have an active participant playing an investigative role. Similarly, a therapeutic act is distinguished from a clinical act by the restriction that they have a continuant playing a therapeutic role.

An echocardiogram diagnostic process can be represented in this way (using the American Heritage Stedman's Medical Dictionary definition):

```

diagnostic-procedure THAT
  hasMethod SOME UltrasoundRecording AND
  investigates SOME
    ( hasConsequence SOME
      ( pathological-disposition THAT
        ro:located_in SOME Heart ) ) AND
  hasOutput SOME Echocardiogram
  
```

The **investigates** relation holds between a diagnostic procedure and either an *etiological agent* or some indication of a **therapeutic-act**. The former is defined in 3.6 along with a *pathological disposition*. An Echocardiogram is a representational artifact, which is defined in the next section. In our example, the thrombolysis was indicated (**hasIndication**) by the result of one or more diagnostic procedures.

3.3 Representational Artifact

In [17], a *representational artifact* is defined as an idea, image, record, or description that refers to (is of or about), or is intended to refer to, some entity or entities external to the representation.

It is the basis for a distinction between information content [5] and the phenomena they represent, or describe. Its use in the CPR ontology is meant to resolve certain ontological inconsistencies [16] that exist in common vocabulary standards such as HL7 RIM that

⁴ a patient's account of a medical history

conflate the two. It denotes CPR content or a component thereof that stand in the **representationOf** relation to some entity.

The primary representational artifact relevant to the domain is the **clinical-artifact** class. It is explicitly differentiated from a representational artifact by the restriction that its instances are composed by a person playing a relevant role in the care process and have a person as the subject of their description. The **subjectOfDescription** relation holds between a clinical artifact and its subject. It is similar to the *SUBJECT RELATIONSHIP CONTEXT* attribute in SNOMED-CT that is used to specify the subject of the *clinical finding* or *procedure* being recorded. Clinical artifacts are also often consumed by (**hasInput**) or produced from (**hasOutput**) a **clinical-act** such as an echocardiogram diagnostic process.

3.4 Clinical Findings

Expanding on the definition from [13], the CPR ontology defines the **clinical-finding** primitive as a **clinical-artifact** that is composed by a clinician, is the **outputOf** a **clinical-act**, and represents a bodily feature. Examples of findings from the case report are the medical history of hypertension (**anamnesis**) and akinesia of the posterior-basal revealed from the previous echocardiogram.

3.5 Surgical Deeds

This CPR ontology distinguishes a **procedure** from a **clinical-act** by the following explicit existential role restrictions: 1) it **actsOn** [8] an organismal continuant [12] or object, as in the case of device insertions; 2) it stands in an *approach site* relation to an immaterial anatomical entity (a cavity, whole, etc.); 3) It stands in a **hasMethod** relation⁵ with the deed [8] or method that achieved it. SNOMED-CT specifies the semantics of the *SURGICAL APPROACH* attribute as the directional, relational, or spatial access to the site of a surgical procedure.

⁵ SNOMED-CT defines a **METHOD** attribute of procedures that relates it to the action being performed to accomplish the procedure.

A **diagnostic-procedure** (such as an echocardiogram) is a primitive procedure that is also a **clinical-investigation-act**.

3.6 Disease Manifestation, Etiology, and Pathophysiology

The concept of a disease as a pathological disposition and its relation to a disease course is introduced in [13]. A disposition is an attribute of an organism in virtue of which it will initiate specific processes when certain conditions are satisfied. Whitbeck's theory [18] of *etiology* and *disease entities* is incorporated in the CPR ontology **pathological-disposition** term.

It is a primitive that is explicitly distinguished from a disposition via the following existential role restrictions: 1) it is located in a patient 2) it stands in the **isConsequenceOf** relation to one or more instances of **etiological-agent**, either directly or via transitive closure of this causal relationship; 3) it is realized⁶ as a pathological process that has, as a component, consistent *physiological* or *anatomical alterations* (**morphological-alteration**). The **isConsequenceOf** relation denotes a causal relationship between entities (occurrent and continuant alike).

3.7 Diagnosis as Clinical Analysis and Hypothesis

Whitbeck provided [19] an ontological basis for clinical diagnoses, how they fit into the terminology of clinical medicine, and their relationship to diseases and etiological agents. The CPR ontology defines a **clinical-diagnosis** as a clinical artifact that is the output of a **clinical-investigation-act**. It stands in the **hypothesizedProblem** relation to a disease that is hypothesized by a clinician to inhere in the patient. Intuitively, this is the recording of the conclusion to an interpretive (scientific) process that investigates problems that require management.

⁶ Currently, a causal, agentive relation is used, however, it is not clear (given [13, 18]) if the relationship between a disease and anatomical alterations is causal or mereological.

3.8 Relationship with Measurement Unit Ontology

The CPR ontology includes a **measurementOf** relation that serves as a bridge between it and the Measurement Unit Ontology (MUO)⁷. It holds between a clinical finding and a **muo:QualityValue**. The MUO defines a *quality value* as the (reified) value of a quality that is related to exactly one unit of measure (via the **muo:measuredIn** relation) and a corresponding scalar quantity (via the **muo:numericalValue** relation).

So, the semantics of the **measurementOf** relation can be captured in the following role chain:

$$\text{subjectOfDescription} \circ \text{bearerOf} \circ \text{measuresQuality}^- \circ \text{measuredIn}^-$$

In this representation, *bearerOf* corresponds to the implicit BFO relationship between an independent continuant and the dependent continuant (quality, function, etc.) that inheres in it and **muo:measuresQuality** is the MUO relation that holds between a unit and the quality it measures.

An alternative approach is to create a new role under **muo:qualityValue** in the role hierarchy that holds between a continuant and the value of a quality associated with the role. So, going back to our medical case, the diastolic and systolic blood pressure values recorded as a result of the physical examination can be captured using a *diastolic* and *systolic* role each of which is subsumed by **muo:qualityValue** and understood to hold between a person and the values of the diastolic and systolic blood pressure qualities, respectively and at the time of measurement.

4 Discussion

In formulating and curating this ontology, there were a number of issues that straddle the line between metaphysics and the philosophy of medicine but are mentioned here.

4.1 Qualities of occurrents

In Sayed's comparison [14] of BFO and DOLCE, the following question was raised:

⁷ http://forge.morfeo-project.org/wiki/en/index.php/Units_of_measurement_ontology

Why can't a heartbeat rate be a quality of its heartbeating [sic] event, given it has no meaning outside of this event?

The consequence of BFO's requirement that qualities can only inhere in continuants is that not all bodily features can have qualities, and thus a concept corresponding to the quality of a heart-beating event or process cannot be straightforwardly represented. A similar example is from the medical case.

Ejection fraction is a measure of the fraction of blood pumped out of the right and left ventricles with each heart beat. Intuitively, it seems inappropriate to consider it a measure of the quality inherent in the heart or in the ventricles but rather as a measurement of the flow of blood through the latter.

4.2 Transitive Causal Chains

Another issue is related to transitive, causal chains between two continuants or where the agent is a process. The CPR ontology re-uses the *ro:agent* in from the Relations Ontology. However, that relation is designed to explicitly rule out the inheritance of agency along causal chains [15] and to ensure that only continuants are agents. This restriction on *ro:agent* in and the fact that a disease is modeled as a continuant presents a challenge and is the reason why the **isConsequenceOf** is needed.

In describing an etiological agent, Whitbeck gives two criteria that govern the choice of the causal factor to be regarded as the etiologic agent. The first is proximity and the second is a preference for a factor which exists in the environment prior to contact with the patient's body and which may then act upon it [18]. She describes the rationale for this second criterion in the way in which we classify diseases in which it is the toxins elaborated during metabolism and growth of the pathogen which cause the damage.

A detailed account of this shortcoming and of the role of causality in an ontological account of clinical medicine is beyond the scope of this paper. However, the interested reader should see [18].

4.3 Signs, Symptoms, and Epistemology

As mentioned earlier, there are issues associated with the semantics of the medical

terms *symptom* and *sign*. In the CPR ontology, they are modeled as representational artifacts and not special kinds of bodily features. It is generally a bad idea to mix ontology and epistemology [10]. The former is about things in reality and the latter is regarding how cognitive subjects come to know the truth about phenomena. Some bodily features can exist before they are perceived by a patient. The patient may subsequently hypothesize that the bodily features are symptoms that indicate a disease. However, they are not considered symptoms initially.

So, there is nothing about a cough (for example) that makes it a symptom besides the way it is first perceived and then reported. Even if a symptom or finding is recorded by a nurse as a result of eliciting anamnesis from the patient, it is the patient that originally presents with it. This is not an ontological distinction and is mostly also the case with signs as well. However, what makes a bodily feature a sign is a little more objective [13]. A sign is typically observed in a **clinical-examination** that is performed by a clinician. Still, the primary distinction between signs, symptoms and other bodily features is with regards to who observed them.

Since this distinction can be captured via the **composedBy** relation between a **clinical-artifact** and a **person** the CPR ontology has two defined classes: **record-of-symptom** and **record-of-sign**. It also has a **self-examination** class that is an action performed by a patient on his or herself to determine the existence of a medical problem.

So, the logical distinction between a symptom and other bodily features is that they are represented by a clinical artifact with provenance indicating authorship by the patient.

5 Conclusion

This paper introduces a minimal, high-level framework that addresses the CPR requirements of a core, common, data dictionary and facilitates the meaningful use of discrete CPR content via biomedical ontologies. It contributes to meaningful use in the following ways: ontology is used to facilitate data that is discrete, precise in its meaning for the purpose of communication and use in different contexts

such as reporting quality measures and clinical care and research information system infrastructure, and as a framework for capturing a large number of clinical conditions.

In using an independent, running example of a medical case, we are able to demonstrate how the CPR ontology acts as a coordinate system that is sufficient to capture the key axes of meaning. As a result, this ontology is able to post-coordinate terms from within a large region of the domain of interest: patient record content. It supports terms that capture the specific meaning of clinical findings, problems, procedures, and treatments.

This ontology has been used as part of the *SemanticDB*⁸ project in the Cleveland Clinic's Heart and Vascular Institute, where the author was a lead software architect. It was also used as part of the development of recent research methods for managing disparate relational, polysomnography data via web-based management [1, 9].

A number of issues are still unresolved and raised as input to the frameworks leveraged by the CPR ontology. The ontology is part of a project hosted on Google Code repository⁹. Any feedback or comments are welcome and can be directed to the associated Google Group¹⁰.

Acknowledgements

The author would like to thank the following people for their substantive input into this ontology: Alan Rector M.D. and Ph.D., Sivaram Arabandi M.D. and M.S., Eugene Blackstone M.D., Barry Smith Ph.D., Michel Dumontier Ph.D., Songmao Zhang Ph.D., GQ Zhang Ph.D., Satya Sahoo Ph.D., Helen Chen Ph.D., and Jos Deroo.

References

1. S. Arabandi, Ogbuji C., Redline S., Chervin R., Boero J., Benca R., and Zhang G.Q. Developing a Sleep Domain Ontology (abstract). American Medical Informatics Association, 2010.
2. D. Blumenthal. Launching HITECH. New England Journal of Medicine, 2010.

⁸ <http://www.w3.org/2001/sw/sweo/public/UseCases/ClevelandClinic/>

⁹ <http://code.google.com/p/cpr-ontology>

¹⁰ <http://groups.google.com/group/cpr-ontology>

3. O. Bodenreider. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*, 67:79, 2008.
4. R.S. Dick, E.B. Steen, and D.E. Detmer. The computer-based patient record: an essential technology for health care. *Natl Academy Pr*, 1997.
5. M. Dumontier and R. Hoehndorf. Realism for scientific ontologies. In *Proceeding of the 2010 conference on Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010)*, pages 387–399. IOS Press, 2010.
6. L. Iuliano, F. Micheletta, A. Napoli, and C. Catalano. Myocardial infarction with normal coronary arteries: a case report and review of the literature. *Journal of Medical Case Reports*, 3:24, 2009.
7. B. Kadry, I.C. Sanderson, and A. Macario. Challenges that limit meaningful use of health information technology. *Current Opinion in Anesthesiology*, 23(2):184, 2010.
8. A. Mori, A. Gangemi, G. Steve, F. Consorti, and E. Galeazzi. An ontological analysis of surgical deeds. *Artificial Intelligence in Medicine*, pages 361–372, 1997.
9. C. Ogbuji, S. Arabandi, S. Zhang, and G.Q. Zhang. Segmenting and merging domain-specific ontology modules for clinical informatics. pages 414–427, 2010.
10. B.S. Olivier Bodenreider and A. Burgun. The ontology-epistemology divide: A case study in medical terminology. In *Formal ontology in information systems: proceedings of the Third International Conference (FOIS-2004)*, page 185. Ios Pr Inc, 2004.
11. A.L. Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 121–128. ACM, 2003.
12. C. Rosse, A. Kumar, J.L.V. Mejino Jr, D.L. Cook, L.T. Detwiler, and B. Smith. A strategy for improving and integrating biomedical ontologies. In *AMIA Annual Symposium Proceedings*, volume 2005, page 639. American Medical Informatics Association, 2005.
13. R.H. Scheuermann, W. Ceusters, and B. Smith. Toward an ontological treatment of disease and diagnosis. *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*, pages 116–120, 2009.
14. A.P. Seyed. BFO/DOLCE Primitive Relation Comparison. 2009.
15. B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A.L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005.
16. B. Smith and W. Ceusters. HL7 RIM: An incoherent standard. *Studies in health technology and informatics*, 124:133–138, 2006.
17. B. Smith, W. Kusnierczyk, D. Schober, and W. Ceusters. Towards a reference terminology for ontology research and development in the biomedical domain. In *Proceedings of KR-MED*, volume 2006, pages 57–65. Citeseer, 2006.
18. C. Whitbeck. Causation in medicine: The disease entity model. *Philosophy of Science*, 44(4):619–637, 1977.
19. C. Whitbeck. What is diagnosis? Some critical reflections. *Theoretical Medicine and Bioethics*, 2(3):319–329, 1981.

ICBO Posters



ICBO

International Conference on Biomedical Ontology

July 28-30, 2011
Buffalo, New York, USA

The Blood Ontology: An Ontology in the Domain of Hematology

Mauricio Barcellos Almeida¹, Anna Barbara de Freitas Carneiro Proietti², Jiye Ai³, Barry Smith⁴

¹Federal University of Minas Gerais, School of Information Science, Belo Horizonte, Brazil

²Hemominas Foundation, Belo Horizonte, Brazil

³University of California, Los Angeles, School of Dentistry and Dental Research Institute, USA

⁴State University of New York at Buffalo, USA

Abstract. Despite the importance of human blood to clinical practice and research, hematology and blood transfusion data remain scattered throughout a range of disparate sources. This lack of systematization concerning the use and definition of terms poses problems for physicians and biomedical professionals. We are introducing here the Blood Ontology, an ongoing initiative designed to serve as a controlled vocabulary for use in organizing information about blood. The paper describes the scope of the Blood Ontology, its stage of development and some of its anticipated uses.

Keywords: hematology, blood transfusion, human blood, human fluids, ontology

1 Background

The biomedical field is vast and complex and the representation of medical facts is also a complex task. The complexity and importance of the medical domain require representation as consistent as that offered by ontologies. The profusion of medical ontologies seen in recent years and the appearance of open data repositories, such as the Open Biomedical Ontologies Foundry [1], can attest to the feasibility of this approach for the life sciences.

Within this context, we introduce the Blood Ontology (BLO), which has been designed to serve as a comprehensive infrastructure resource allowing for the exploration of information relevant to scientific research and human blood manipulation. The BLO is part of a long-term ongoing knowledge management project in the field of hematology and blood transfusion, structured according to three main axes: i) knowledge organization based on ontological principles; ii) knowledge acquisition from experts and texts; and iii) visualization tools. In this paper, we describe the BLO and its current stage of development.

2 Methods

The BLO consists of a set of co-related ontologies, each one addressing a group of relevant issues in the field of hematology and blood transfusion. These sub-ontologies are: i) BLO-Core, an ontology of hematological essentials; ii) BLO-Management, an ontology for the management of blood-related processes; iii) BLO-Products, an ontology representing the products resulting from blood manipulation; and iv) BLO-Administrative, an ontology for regulatory documents.

In the current stage, the main activities being performed for BLO-Core are: i) the collection of terms from the domain of hematology; ii) reuse of data available in other ontologies; iii) organization of a hierarchy and prospective studies of relationships.

Knowledge acquisition is being undertaken by experts in biology and medicine, who are members of the Hemominas Foundation¹, the second largest Brazilian blood bank. We have been using the following methods: i) interviews oriented according to forms created in Protégé-Frames, which are based on the Ontology for General Medical Science (OGMS) [2]; ii) validation of

¹ <http://www.hemominas.mg.gov.br/>

knowledge acquired using a semantic wiki; iii) translation of validated terms from the wiki to Protégé-OWL.

With the aim of fostering interoperability among ontologies, the BLO relies on well-consolidated initiatives, namely those pertaining to the OBO Foundry framework [4]. Within the OBO scope, important initiatives are the Gene Ontology [3], the Protein Ontology [4], the Cell-Type Ontology [5], to mention but a few. BLO also relies on the foundational grounds of the Basic Formal Ontology [6], an upper-level ontology created to support scientific research. In order to gather such ontologies we have adopted the experimental approach called Minimal Information to Reference External Ontology Terms (MIREOT) [7], which was developed as part of the Ontology for Biomedical Investigations project [8].

3 Results

This section describes the current stage of development of each BLO sub-ontology, as well as the planned scope of future developments.

The BLO-Core focuses on physiological aspects of blood and presents the basic information required to work on hematological research and practice. The ontology provides the essentials of the chemical constituents and of the molecular, immunologic and cellular basis of blood, as well as blood disorders and transplantation. Currently, more than eight hundred terms have been defined and incorporated into the Core subset. (Preliminary results are available at: http://mbaserver.eci.ufmg.br/BLO-wiki/index.php/BLO_Core)

The ontology named BLO-Management covers the relevant processes involved in blood manipulation and related services. Blood manipulation involves primarily the following activities: i) quality management; ii) blood utilization management; iii) donor selection; iv) blood collection; v) control of transfusions; vi) control of apheresis; vii) blood testing. In turn, these activities involve a range of processes that were considered in the design of BLO-Management. This branch of the BLO is under development involving natural language processing techniques applied to a

set of documents from the American Association of Blood Banks.

The BLO-Products ontology is aimed at facing the challenges created by a multitude of possible “product” derivatives of blood manipulation managed on a worldwide scale. It is mainly based on studies about ISBT-128 [9], an internationally defined labeling system. ISBT-128 standardizes a bar-coded symbology for blood products, allowing them to be read at blood banks and transfusion services around the world.

A standard as ISBT has proven to be of practical importance, but it allows for the ambiguities that are common in a natural language. This branch of the BLO is under development and the results are also partial. The research involves the evaluation of rules used to create IBST terms and descriptors in order to check ontological decisions underlying that standard.

The BLO-Administrative ontology is aimed at covering the issues related to the official documentation of interest to blood banks and transfusion services. By “documentation”, we mean policies, documents from regulatory agencies, professional class associations, law, regulations, officially recognized classification systems, and standards. BLO-Administrative is being designed in line with the Information Artifact Ontology (IAO) [10]. The results here are partial and concern attempts to create an additional characterization for documents based on a pragmatic approach. Examples of the kinds of documents being evaluated are: blood donation orders, consent letters, quality requirements, to mention but a few.

4 Discussion

The BLO serves several purposes, for example: as the core vocabulary for the development of interoperable systems, as a base for computational inferences, as a knowledge base for educational purposes, as a tool to aid in information for diagnosis.

The importance of a diagnosis based on the components of blood resides in the fact that blood cells are accessible indicators of disturbances in their organs of origin. During illness, abnormalities can develop in any of the cells in the blood and their detection may

aid in diagnosis, as well as in the care of patients.

The BLO also considers the broader perspective of human fluids [11]. It is worth mentioning the collaboration between the BLO and the Saliva Ontology (SALO, <http://www.skb.ucla.edu/SALO/>). SALO is a consensus-based controlled vocabulary of terms and relationships related to the salivaomics domain [12]. It relies on research on salivary diagnostic technologies being developed by the UCLA Salivaomics Research Group.

The BLO is being specialized, for example, for use by the Hemominas Foundation research group, which focuses mainly on the studies of blood transmitted diseases (HIV-1/2, hepatitis B and C, among others), hemophilias, Von-Willebrand disease, Sickle Cell Anemia and Human T-cell Lymphotropic Virus.

These initiatives have been designed to be aligned with the Infectious Disease Ontology (IDO, <http://www.infectiousdiseaseontology.org/>), which is an initiative that gathers together a set of ontologies covering the infectious disease domain.

5 Conclusion

In this paper, we presented the BLO as an ongoing initiative, developed with the aim of facilitating the access to, use and analysis of data on blood. The medical field is a broad and highly developed system to which both medical science and clinical practice contribute. The importance of the BLO resides in the realization that, despite all the advances of recent years, many medical processes pertaining to human blood are still not fully understood.

Acknowledgments

This work is partially supported by *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG), *Belo Horizonte*, MG, Brazil.

References

1. Smith et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, <http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html> (2007).
2. R. Scheuermann, W. Ceusters, B. Smith. Toward an Ontological Treatment of Disease and Diagnosis, http://ontology.buffalo.edu/medo/Disease_and_Diagnosis.pdf (2009).
3. Gene Ontology (GO) Consortium, Gene Ontology, <http://www.geneontology.org/> (2003).
4. A. Natale et al., Framework for a Protein Ontology, Proceedings of the First International Workshop on Text Mining in Bioinformatics (2006).
5. J. Bard, S. Y. Rhee, M. Ashburner, An ontology for cell types, *Genome Biology*, 6 (2) (2005), R21.
6. P. Grenon, B. Smith, L. Goldberg, Biodynamic Ontology: Applying BFO in the Biomedical Domain, in: D. M. Pisanelli (Ed.), *Ontologies in Medicine*, IOS, Amsterdam, 2004, pp. 20–38.
7. M. Courtot et al., MIREOT: the Minimum Information to Reference an External Ontology Term, <http://precedings.nature.com/documents/3574/version/1> (2009).
8. R.R. Brinkman et al. Modeling biomedical experimental processes with OBI. *Journal of Biomedical Semantics* 1Supl. (2010) 1-10.
9. International Council for Commonality in Blood Banking Automation (ICCBBA), ISBT-128: Standard Terminology for Blood, Cellular Therapy, and Tissue Product Descriptions, <http://iccbba.org/>.
10. A. Ruttenberg et al., From Basic Formal Ontology to the Information Artifact Ontology, <http://icbo.buffalo.edu/presentations/Ruttenberg.pdf> (2009).
11. W. Yan et al., Systematic comparison of the human saliva and plasma proteomes. *Proteomics Clinical Applications* 3 (2009)116–134.
12. J. Ai, B. Smith, D.T. Wong, Saliva Ontology: An ontology-based framework for a Salivaomics Knowledge Base, <http://www.biomedcentral.com/1471-2105/11/302/abstract> (2010).

Employing Reasoning within the Phenoscape Knowledgebase

James P. Balhoff^{1,2}, Wasila M. Dahdul³, Hilmar Lapp¹, Peter E. Midford¹,
Todd J. Vision^{1,2}, Monte Westerfield⁴, Paula M. Mabee³

¹National Evolutionary Synthesis Center, Durham, NC, USA

²University of North Carolina, Chapel Hill, NC, USA

³University of South Dakota, Vermillion, SD, USA

⁴University of Oregon, Eugene, OR, USA

The Phenoscape project (<http://phenoscape.org>) links evolutionary phenotype descriptions to model organism phenotypes via ontological annotation of comparative data, with the ultimate goal of generating hypotheses about the genes involved in evolutionary phenotype transitions. To date, we have created ontological phenotype descriptions for over 11,000 evolutionary character states linked to 2500 ostariophysan fish taxa. We combined these annotations with mutant phenotype annotations derived from the Zebrafish Information Network (ZFIN), along with shared anatomy, quality, and taxonomy ontologies, into the Phenoscape Knowledgebase. The utility of ontological annotations is derived from the use of shared identifiers for concepts and the explicit computable semantics of those concepts provided by the ontologies. By building on a semantic data store and reasoning system (OBD, the Ontology-Based Database), the Phenoscape Knowledgebase allows users of its web interface to exploit the rich semantics of phenotypic annotations.

Because phenotypes are expressed as complex intersections of anatomy and quality (and often other) ontology terms, the Knowledgebase relies on automatic classification by the OBD reasoner to infer subsumption of phenotypes by other phenotypes as well as placement within the core ontology hierarchies. Class subsumption, transitive properties, and property chains are all employed in providing results within the

Knowledgebase web query interface. For example, users are able to specify a search for all phenotype annotations involving the "shape of any part of the head, within the taxon Cypriniformes". By default, queries within higher taxa (e.g. an order such as Cypriniformes) return annotations to any of their component subtaxa. However, via an option in the query interface, users can make use of additional semantics to automatically infer that phenotypes annotated to a higher taxon apply to all its members.

The OBD reasoner provides an interface for custom rules to be incorporated into the reasoning framework. Using this interface, the Phenoscape Knowledgebase allows automatic inference of absence phenotypes for structures which develop from (as stated in the anatomy ontology) structures which have been asserted as absent within a given taxon. Using the reasoning framework and rule interface allows domain-specific assumptions such as these to be stated explicitly and not buried in application code.

Phenoscape is continuing to address other reasoning challenges. We are developing a logical framework for homology assertions which will allow proposed homologies to be correctly propagated in the context of anatomy ontologies and phenotype annotations. Additionally, we are exploring methods for incorporating taxon-specific anatomical relations into our anatomy ontology which can facilitate queries within specific taxonomic contexts.

Waiting for a Robust Disease Ontology: A Merger of OMIM and MeSH as a Practical Interim Solution

Susan M Bello¹, Allan Peter Davis², Thomas C Wiegers², Mary E Dolan¹, Cynthia Smith¹,
Joel Richardson¹, Judith Blake¹, Carolyn Mattingly², Janan T Eppig¹

¹Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME

²Comparative Toxicogenomics Database, The Mount Desert Island Biological Laboratory, Salisbury Cove, ME

Abstract. The curation of human diseases suffers from the lack of a robust, publicly available disease ontology. While waiting for such an ontology, bioinformatics resources associating genetic and genomic data with human diseases need an interim solution. The Comparative Toxicogenomics Database (CTD) produced a practical, structured vocabulary by curating the association of two subsections of the National Library of Medicine's Medical Subject Headings (MeSH) with Online Mendelian Inheritance in Man (OMIM). The MeSH subsections provide hierarchical access to broad disease terms. OMIM provides detailed disease descriptions and links to associated human genes and mutations. Both resources are freely available and familiar to the research community. Mouse Genome Informatics (MGI) reviewed and modified this vocabulary to adapt it for curation of mouse models of human disease. Here we describe the merged vocabulary and discuss the strengths and weaknesses of this approach. CTD's merged MeSH – OMIM vocabulary can be accessed at <http://ctd.mdibl.org/voc.go?type=disease>.

Introduction

The need to curate disease-related data is a pressing one for many databases. There is a growing pool of available experimental data and increasing pressure from funding agencies to make clear connections between disease related data from model organisms and the human diseases they reference. There are a number of disease vocabularies and ontologies that may be used for the purpose of annotating disease models each with its own advantages and disadvantages [1]. At Mouse Genome Informatics (MGI) [2] disease annotations are currently made to OMIM [3]. OMIM was initially selected based on the presence of detailed disease descriptions, links between disease records and human genes, and familiarity to biomedical researchers. However, the absence of hierarchical structure in OMIM and the absence of generic disease classes, such as Parkinson Disease, have resulted in a growing collection of mouse models of human disease that cannot be annotated using OMIM since these mice are only described in terms of a generic form of the disease.

Accordingly, MGI sought to identify a disease ontology or vocabulary to better annotate mouse models of human disease. Criteria for selecting a disease ontology have been published before [1,4]; however, which criteria are considered and how much weight is put on the criteria varies depending on the perspective of the end user. The criteria considered here include several of those described by Bodenreider and Burgen [1] (coverage of diseases, regular maintenance, support for reasoning, open availability). Additional criteria included; stability of the vocabulary, percentage of terms with definitions, inclusion of synonyms and familiarity of the vocabulary to the user community. A final and necessary consideration was presence of links to OMIM.

Several of these additional criteria are generally applicable to any ontology selection process for use in annotation. A stable ontology avoids the need for extensive and repeated re-curation of data. Deep synonym coverage allows for easier identification of diseases from the literature and for more effective searching of the data by users. Definitions provide a description of the disease

to aid in understanding of the disease term and provide a basis for comparison to the model. Familiarity of the user community improves the likelihood that users will readily find the disease(s) for which they are searching. However, the inclusion of links to OMIM was important both for the general value of the OMIM text resources to MGI users and to efficiently use existing disease model annotations in MGI.

Of the existing disease ontologies and vocabularies, two were identified as containing at least some links to OMIM; The Disease Ontology (DO) [5] and the Comparative Toxicogenomics Database's (CTD) combined MeSH and OMIM disease vocabulary [6]. While the DO may grow into a better long-term solution, it is, as of now, not nearly mature or robust enough to be useful for curating disease data. The DO is being extensively revised (negatively impacting stability), only 11% of the terms have definitions (as of 6/21/10), and while OMIM ids are being added many are still missing and there is uneven mapping of OMIM diseases within the DO (Drs. Lynn Schriml and Warren Kibbe, personal communication). Therefore we undertook an extensive review of the CTD merged disease vocabulary.

1 OMIM - MeSH Combined Vocabulary

CTD created, implemented and maintains the merged OMIM-MeSH vocabulary (manuscript in preparation). Two subsections of MeSH were used to create the vocabulary: Diseases [C] and Mental Disorders [F03]. OMIM terms were limited to those with a known gene locus. To merge the vocabularies, all selected OMIM terms were mapped manually to MeSH terms based on semantic similarity (i.e. OMIM 101000, NEUROFIBROMATOSIS, TYPE II, was merged with MeSH D016518, Neurofibromatosis 2; OMIM 125850) or symptom matching for OMIM terms lacking a semantic match. The use of symptoms for mappings allows for rapid and consistent mapping but, curation issues can result and this is discussed more fully below.

Of the 7052 OMIM phenotype or disease records [3], 4365 were associated with a gene map locus. CTD had mapped 4049 of this set to one or more MeSH terms[6]. The 316

unmapped terms were largely records for phenotypic variation (e.g., Hair Morphology 2, OMIM:139450; ABO Blood Group, OMIM:110300). CTD loads the vocabularies on a monthly basis and any discrepancies are identified and curated. The merged vocabulary can be accessed at:

<http://ctd.mdibl.org/voc.go?type=disease>.

2 MGI Review

The CTD disease vocabulary was reviewed to determine the extent of coverage of OMIM terms in use by MGI. The first review (conducted in June 2010) identified 347 OMIM terms in MGI but absent from the CTD disease vocabulary. 259 of these were in CTD's unmapped set. The remaining 88 terms were either new OMIM terms or OMIM terms without a gene map locus. A second review (conducted in August 2010) identified 212 terms in MGI but absent from the CTD disease vocabulary. 37 were repeats from the first review, 90 were new OMIM terms. 85 were existing OMIM terms without a gene map locus. All unmapped OMIM terms were then examined and either mapped to appropriate MeSH terms or added to the unmapped term set.

The reviews also revealed a difference in the desired level of granularity of the vocabulary between CTD and MGI. CTD merged many gene specific OMIM disease terms with the general disease term. For example, AGAMMAGLOBULINEMIA 1 (OMIM 601495) caused by a mutation in *IGHM* and AGAMMAGLOBULINEMIA 6 (OMIM 612692), caused by a mutation in *CD79B*, were both merged with MeSH D000361, Agammaglobulinemia. MGI would prefer both OMIM terms be made children of the MeSH to allow for distinction between orthologous and non-orthologous mouse models. Therefore the CTD OMIM to MeSH mappings were reviewed to identify all instances where an additional level of granularity was desired. The mappings have been internally annotated to record such cases. With these modifications the CTD vocabulary includes terms for all of MGI's existing mouse models of human diseases, allows for annotation of models that cannot be

annotated using OMIM and will improve user access to disease model annotations.

3 Symptom Based Mapping

Mapping of OMIM terms based on disease symptoms has consequences. That a disease produces a symptom in an organ or tissue does not necessarily mean that the disease is a disease of that organ or tissue. For example, in MeSH, albinism is a child of eye diseases and pigmentation diseases, while experts would agree that albinism is a pigmentation disease, many would not consider it an eye disease. However, symptom based mapping can also explain and illuminate a disease. For example, mapping the OMIM term RIDDLE SYNDROME (OMIM 611943) to the MeSH terms for its symptoms (immune deficiency syndromes, learning disorders, and facies) provides insights into the disease. Unfortunately, some symptom descriptions may lead to erroneous mappings if the mapping is not constructed or reviewed by a clinician. Symptoms described as being “like” some other disease or syndrome, may be semantically, yet erroneously, mapped to that disease. For example, patients with Lujan-Fryns Syndrome are described as having “Marfanoid habitus”, a term seemingly related to the term ‘Marfan’ but whose definition is not related to Marfan Syndrome. The symptom based association results in a mapping of Lujan-Fryns Syndrome to Marfan Syndrome, which is incorrect. These kinds of situations require experts in disease phenotypes to identify, review and curate. Such clinical experts must be an integral part of any disease ontology development effort.

4 The Future

CTD's merged vocabulary is an interim solution to a pressing curation need. Both MGI and CTD plan to migrate annotations to a comprehensive disease ontology, once one is mature and ready for broad use. The merged vocabulary should inform development of a new disease ontology and incorporation of OMIM and MeSH identifiers into developing disease ontologies will greatly aid in future

migration of existing annotations to a new ontology.

Acknowledgments

CTD is supported by ES014065 from NIEHS/NLM, ES014065-04S1 from NIEHS, and P20RR016463 from NCRR. MGI is supported by HG000330 from NHGRI/NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Bodenreider, O., and Burgun, A.: Towards desiderata for an ontology of diseases for the annotation of biological datasets. ICBO 39-42 (2009).
2. Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., Eppig, J.T. and the Mouse Genome Database Group.: The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* 39(suppl 1), D842--D848 (2011)
3. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), <http://omim.org/> (2/9/2011)
4. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25(11):1251-5 (2007).
5. Osborne, J.D., Flatow, J., Holko, M., Lin, S.M., Kibbe, W.A., Zhu, L.J., Danila, M.I., Feng, G., Chisholm, R.L.: Annotating the human genome with Disease Ontology. *B.M.C. Genomics* 10, S1:S6 (2009)
6. Davis, A.P., King, B.L., Mockus, S., Murphy, C.G., Saraceni-Richards, C., Rosenstein, M., Wieggers, T., Mattingly, C.J.: The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.* 39(Database issue), D1067--72 (2011)
7. Medical Subject Headings, <http://www.nlm.nih.gov/mesh/>

Developing a Reagent Application Ontology within the OBO Foundry Framework

Matthew H. Brush¹, Nicole Vasilevsky¹, Carlo Torniai¹,
Tenille Johnson², Chris Shaffer¹, Melissa A. Haendel¹

¹Oregon Health and Science University, Portland, OR, USA; ²Harvard University, Cambridge, MA, USA

1 Introduction

In light of the increasing complexity and cost of biomedical research, the ability to find and re-use research resources has emerged as an important challenge for information technologies to address. Among the most expensive resources to produce, and challenging to discover, are the material reagents employed in performing biomedical experiments. The eagle-i Consortium (www.eagle-i.org/home) represents an open and collaborative effort to develop ontology-driven applications aimed at cataloging biomedical research resources, including reagents, and making them available through a semantic search portal. Here, we describe the development of a reagent ontology module within our larger eagle-i ontology, that uses Open Biomedical Ontologies (OBO) Foundry [1] best practices to facilitate interoperability and extend external ontology efforts.

As part of an ‘application ontology’, key goals for the eagle-i reagent module were to represent real data collected from research laboratories, and to drive the user-interface (UI) and application logic of data collection and search tools. Reagent modeling was informed by analysis of data collected from participating laboratories, whereby seven top-level reagent categories were identified. These included ‘Antibody,’ ‘Cell Line,’ ‘Chemical Reagent,’ ‘Construct,’ ‘Nucleic Acid Reagent,’ ‘Protein Reagent,’ and ‘Reagent Library’ classes. An equally important goal in developing the reagent module was to facilitate interoperability and linking of data across systems, such that reagents described in eagle-i could be linked to experiments, protocols, publications, or data sets cataloged in external sources. This Linked Open Data (LOD) [2] approach can add significant value to data repositories, particularly when

their underlying ontology models are logically integrated to allow computational reasoning across linked bodies of data. This approach serves not only to enhance the performance of the eagle-i search portal by exploding the information linked to cataloged resources, but can also support a range of external applications that can be fed by the data housed in the eagle-i repository. To promote interoperability and publication of LOD, we looked to OBO Foundry Library as set of orthogonal ontologies with which to align our efforts. The OBO Foundry provides an evolving set of shared principles for ontology development, and houses a collection of ‘reference ontologies’ designed according to these standards. In developing our reagent ontology module, OBO principles and reference ontologies were consulted and reused as described below.

2 The Ontological Nature of Reagents

The term ‘reagent’ means different things to different people. For example, in chemistry a reagent is considered an inert substance that catalyzes a reaction, while in biomedical research, reagents span a broader range of granularities, and might be any material entity input into an experiment (drugs, antibodies, cell lines, etc). In this sense, reagents are defined by their playing a role a scientific experiment. For instance, antibodies exist naturally in organisms where they serve to fight infection, but those applied to detect some analyte in an experimental setting are considered reagents. Similarly, many chemical compounds are produced naturally through biological processes (antibiotics, toxins, etc), but qualify as reagents only if used in a research investigation. To capture this idea, we have defined reagents as “material entities used in an experimental process to detect, measure,

examine, or produce other substances”. Accordingly, classification of a material entity as a reagent depends not on some shared physical attribute, but rather on its use in a particular context (biomedical experimentation) and for a specific purpose (to generate data or other materials for experimentation). Therefore, rather than assert that material entities are members of a reagent class in our ontology, they are classified with other entities that share common inherent physical features, and inferred to be reagents based on an axiom asserting their role in scientific experimentation. For example, a ‘plasmid’ is classified as a subtype of ‘double-stranded DNA’, and described by an axiom that indicates it to play a ‘reagent role’. As described below, the OWL-DL language in which our ontology is written offers mechanisms for using such axioms to generate a unified hierarchy of reagents that meets our application needs.

3 Preliminary Landscape Analysis

Two key principles advocated by the OBO Foundry are *interoperability*, whereby existing ontologies and classes are re-used whenever possible, and *orthogonality*, whereby modeling of a particular domain converges upon a single reference ontology. In keeping with these principles, we performed a preliminary analysis of existing ontological representations of reagents to determine whether any might offer a single, comprehensive model that could be adapted to meet our application-specific needs. An examination of the 266 ontologies cataloged by the National Center for Biomedical Ontology (NCBO) Biportal [3] revealed very limited modeling of reagents. A ‘reagent’ class appeared in fewer than ten ontologies, and none of these offered sufficient logical descriptions or classification of reagent subtypes for our application needs. What sparse modeling had been done was inconsistent, with ‘reagent’ classes being modeled alternately as ‘material entities’, ‘roles’, and ‘features’ in different ontologies. This modeling was also restricted in scope, often describing only those reagents specific for a particular domain, application, or granularity. Therefore, convinced that no existing resource offered a suitable, comprehensive, and broadly applicable model of

reagents for our needs or that of the community at large, we set out to construct one.

4 Modeling Approach

Having found no suitable representation of reagents among existing ontologies, our landscape analysis was extended with two goals in mind: (1) to identify “source” ontologies from which to re-use individual classes representing reagent types (ensuring interoperability); and (2) to identify an appropriate “home” ontology in which to re-use these classes to construct our reagent model (ensuring orthogonality). Below, we describe our plans for identifying source and home ontologies, and their roles in developing a stand-alone reagent ontology module that meets eagle-i application needs.

4.1 Source Ontologies for Reagents

Reagents are represented across many levels of granularity – from small molecules such as drugs and chemicals, to biological macromolecules such as proteins or DNA constructs, to cells and cell lines, to libraries comprised of large collections of peptides or genomic clones. Many of these reagent types are represented across the OBO library of ontologies, but are not defined therein as reagents. For example, proteins are modeled in the Protein Ontology (PRO) [4], chemicals are represented in the Chemical Entities of Biological Interest Ontology (ChEBI) [5], and many types of nucleic acids are modeled in both ChEBI and the Sequence Ontology (SO) [6]. When materials representing reagents in our seven top-level categories fall within the scope of existing ontologies, relevant classes from these sources will be imported into our home ontology, where they will be extended to model reagents. For example, because all instances of eagle-i ‘Construct’ reagents are types of ‘ChEBI: nucleic acid > DNA > double-stranded DNA’, these classes will be imported from ChEBI into our home ontology. Here, construct reagents and their more specific subtypes (e.g. ‘plasmid’, ‘viral plasmid’, ‘retroviral plasmid’), which lie outside the scope of ChEBI, will be modeled beneath the ChEBI ‘double-stranded DNA’ class.

4.2 A Home Ontology for Modeling Reagents

As reagent classes are defined by their participation in experimental processes, we determined that the process-oriented Ontology for Biomedical Investigations (OBI) [7] would serve as a suitable “home” for our modeling efforts. OBI is an actively developed OBO foundry candidate ontology driven by representatives from over 20 research communities, which describes all phases of biomedical investigations. OBI also models the material entities that participate in these processes, including reagents, instruments, specimens, and agents. Furthermore, OBI has defined a mechanism for importing (re-using) terms from external sources, called MIREOT (Minimum Information to Reference an External Ontology Term) [8]. Thus, classes representing reagents that are implemented in various “source” ontologies can be imported into a single “home” ontology (OBI) using the MIREOT principle.

Once all relevant classes are assembled in OBI, we will be able to infer a stand-alone hierarchy of reagents sufficient to drive our suite of eagle-i applications. This will be achieved by attaching logical axioms to all classes representing reagents asserting that they *necessarily* bear a reagent role. A ‘Reagent’ equivalent class can then be created, and defined with a *necessary and sufficient* axiom equating it to any ‘material entity’ that has a ‘reagent role’. An OWL-DL reasoner can then generate an inferred ‘Reagent’ hierarchy that unites all reagents in OBI. This will provide a single reagent module, encoded in OWL/RDF, that can be imported into our eagle-i ontology to drive application functionality, while also providing interoperability with external efforts that point to shared OBO classes we used in our model.

5 Conclusions

Our approach for application ontology development aims to balance practical project requirements with OBO best-practices that support interoperability and orthogonality. While this approach creates some initial overhead in the form of a landscape analysis and technical hurdles to re-using of existing

terminologies, it offers many long-term benefits for both application developers and the community at large. For eagle-i, our reagent modeling efforts have been enhanced in many ways by aligning with OBO principles. First, we have benefited from the collaborative nature of ontology development, applying feedback and resources from Foundry members toward improving our reagent model. Second, we are contributing back to the community by extending several Foundry ontologies to include classes and properties relevant to our domain of interest. Third, we are enabling our application to access and use data from external systems built on models that comply with OBO foundry principles. Finally, we hope to provide a reagent ontology module suitable for re-use in the community, which will continue to add value to eagle-i by expanding data available to the system.

References

1. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007;25(11):1251-5.
2. Bizer C HT, Berners-Lee, T. Linked data-the story so far. *Int. J. Semantic Web Inf. Syst.* 2009;5:1-22.
3. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *Omics* 2006;10(2):185-98.
4. Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, et al. The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res*;39(Database issue):D539-45.
5. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;36(Database issue):D344-50.
6. Eilbeck K, Lewis SE, Mungall CJ, et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* 2005;6(5):R44.
7. Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, et al. Modeling biomedical experimental processes with OBI. *J Biomed Semantics*;1 Suppl 1:S7.
8. Courtot M GF, Lister AL, Malone J, Schober D, Brinkman RR, Ruttenberg A. MIREOT: the minimum information to reference an external ontology term. *Nature Proceedings* 2009.

The Ontology of Microbial Phenotypes (OMP): A Precomposed Ontology Based on Cross Products from Multiple External Ontologies that is Used for Guiding Microbial Phenotype Annotation

Marcus Chibucos¹, Adrienne Zweifel², Deborah Siegele², Peter Uetz^{3,4}, Michelle Giglio¹, James Hu²

¹Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA

²Texas A&M University, College Station, TX, USA

³University of Delaware, Newark, DE, USA

⁴Proteros Biostructures, Gaithersburg, MD, USA and Martinsried, Germany

Abstract. The Ontology of Microbial Phenotypes (OMP) is being developed to standardize capture of phenotypic information, including both processes and physical characteristics, from microbes. The OMP team comprises ontologists, microbiologists, and annotators, and ontology development is being performed in conjunction with the development of a wiki designed for annotation capture. Term development is being guided by following, to as great an extent as possible, the structure of existing ontologies. All OMP terms have Aristotelian definitions, and, when appropriate, they have genus-differentia cross products composed of terms from external ontologies. Initially, OMP is being used to annotate the prokaryotic model organism *Escherichia coli*. Eventually we anticipate that diverse user groups will employ OMP for standardized annotation of various microbial phenotypes, much in the same way that the Gene Ontology has standardized the annotation of gene products. Definitions of phenotypes and links to the original literature will facilitate the experimental characterization of phenotypes.

Keywords: annotation capture, bacterial phenotype, Aristotelian definition, cross product, *Escherichia coli*, GONUTS, microbe, microbial phenotype, phenotype annotation, wiki.

1 Introduction

Microbial Phenotypes and the Need for an Ontology

A phenotype is the expression of a genotype (i.e. the full genetic complement of an organism) in a given environment. For example, eye color, number of seeds per pod, and coat color are phenotypic traits that can be observed in flies, lupines, and ponies, respectively. Within an individual organism, both changes in genetic makeup, such as from bacterial conjugation, and variation in gene expression can result in different phenotypes under similar environmental conditions. Conversely, environmental variation can lead to different outcomes for genetically identical organisms, through variable gene expression. Myriad genetically and taxonomically diverse

microbes exhibit countless variability in their morphological and physiological traits, both within and among species. Oftentimes these result in unique and exquisite manifestations, such as the symbiosis between the bioluminescent *Vibrio fischeri* bacterium and its squid host *Euprymna scolopes*. Characterization of phenotypes is critically important for medical microbial identification, and many unique biotechnological applications of microbes are rooted in phenotypes. Genetic manipulation with associated phenotypic characterization remains an important tool for determining protein function in microorganisms amenable to manipulation, such as *Escherichia coli*. To facilitate research in all of these areas, we are developing the Ontology of Microbial Phenotypes to allow for standardized capture of essential phenotypic information.

2 Previous Work

Manual versus Cross Product Terms

Previously, we explored two parallel approaches for building the Ontology of Microbial Phenotypes (OMP) [1] using the ontology editor OBO-Edit [2] and a custom script. We read 100 papers involving metabolic phenotypes and from these we identified 40 microbial phenotypes. We manually created an ontology comprising five super classes to represent those phenotypes. Separately, we created an ontology of cross products between selected PATO [3] terms and two nodes from the Gene Ontology [4], *cellular carbohydrate metabolic process* (GO:0044262) and *cellular amino acid metabolic process* (GO:0006520), which encompassed those phenotypes. Both approaches had advantages. Creating automated cross products was faster; manual term generation often reflected better how a biologist or annotator might think of a phenotype. Both highlighted the importance of creating synonyms and manually wording English definitions in a colloquial syntax, while adhering to strict equivalence guidelines.

Structure of the Ontology of Microbial Phenotypes

The first version of the Ontology of Microbial Phenotypes comprising 252 terms was released in June, 2011, and can be downloaded from SourceForge [5]. As its root class, OMP has *microbial phenotype* (OMP:0000000), defined as “the manifestation of a microbe’s genotype in an environment.” Descended from the root are terms that describe various attributes of microbial phenotypes including: cell arrangement, cell staining, cellular development, cellular morphology, metabolism, motility, multiorganism interactions, and response to stimulus.

Physical Objects and Processes. Phenotypes described by OMP include both physical objects and processes; descendants of the root *microbial phenotype* address both morphological and physiological traits. For example, *absence of flagellum* (OMP:0000030) is a descendant of *cellular morphology phenotype* (OMP:0000071), which is defined as

“a microbial phenotype, where the trait in question is the form and structure of the cell.” However, all of the terms that describe the process of motility are descended from a distinct class *motility phenotype* (OMP:0000001), defined as “a microbial phenotype where the trait in question is the self-propelled movement of a cell from one location to another.”

Relative versus Absolute Phenotypes.

Some phenotypes described by OMP are relative. A process might be altered relative to how a typical organism performs that process, or an organism might not possess a cell part that it would typically possess. For example, a type of microbe that usually possesses a flagellum, but which does not possess a flagellum due to a mutation, is described by the relative term *loss of flagellum* (OMP:0000032). In contrast, some OMP terms describe non-relative characteristics that are inherent in an organism. For example, a type of microbe that does not naturally produce a flagellum would be described by *non-flagellated* (OMP:0000019).

Aristotelian Definitions. All terms have Aristotelian (genus-differentia) definitions of the form “B is an A that C’s.” All terms composed with cross products have cross product definitions, which describe a quality that inheres in a thing, whether a process or an object [6]. For example, the OMP term *abolished motility* (OMP:0000044) has the cross product definition “abolished inheres in cell motility” constructed from the PATO term *abolished* (PATO:0001508) and the GO term *cell motility* (GO:0048870).

3 Wiki for Annotation Capture

We are implementing a wiki modeled on Gene Ontology Normal Usage Tracking System (GONUTS) [7], our wiki for Gene Ontology, for exploring the ontology, adding usage notes to terms, and making phenotype annotations. We use a custom table-editing extension to provide structured data entry and data-mining capabilities. We envision annotation of different types of entities: taxonomic entities such as species and strains, mutant phenotypes, and phenotype predictions.

4 Downloads and Becoming Involved

Two versions of the ontology are available for download; one contains cross products from other ontologies, and the other is a streamlined version with no cross products. Both versions have database cross references, where appropriate. The OMP team welcomes community involvement in the development and application of OMP. We maintain a wiki [8] to facilitate discussion of ontology and annotation related issues, and we have a Source Forge tracker [9] for term requests.

Acknowledgments

We would like to thank NIH/NIGMS for generous funding of this project (R01 GM089636).

References

1. Giglio, M., Mungall, C., Uetz, P., Yin, L., Goll, J., Siegele, D., Chibucos, M., Hu, J.: Development of an Ontology of Microbial Phenotypes (OMP). *Nature Precedings*. (2009)
2. OBO-Edit, <http://oboedit.org>
3. Phenotypic Quality Ontology, http://obofoundry.org/wiki/index.php/PATO:Main_Page
4. The Gene Ontology, <http://www.geneontology.org>
5. SCM Repositories – microphenotypes, <http://microphenotypes.svn.sourceforge.net>
6. Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E., Ashburner, M.: Integrating Phenotype Ontologies across Multiple Species. *Genome Biology*. 11:R2 (2010)
7. Gene Ontology Normal Usage Tracking System (GONUTS), http://gowiki.tamu.edu/wiki/index.php/Main_Page
8. Ontology of Microbial Phenotypes wiki, <http://microbialphenotypes.org/wiki/>.
9. SourceForge site for the Ontology of Microbial Phenotypes, <http://sourceforge.net/projects/microphenotypes>

Towards an Adverse Event Reporting Ontology

Mélanie Courtot¹, Ryan R. Brinkman^{1,2}

¹BC Cancer Agency, Vancouver, BC, Canada

²Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

Reports of adverse events that occur during clinical trials help identify issues with treatment safety and efficacy, and allow for better education of health practitioners and the general public, ultimately increasing patient safety. However, current methods used for spontaneous adverse events reporting are not sufficient, mitigating their usefulness. The Brighton Collaboration is a global network of world-renowned experts providing high quality vaccine safety information. Based on these guidelines, each adverse event following immunization (AEFI) will be decomposed into its constitutive elements (e.g., motor manifestations occurring during seizure), and will be encoded into an ontology. Ontologies are formal representations of knowledge using some well-defined logic. They can streamline the process of integrating, accessing and querying data by providing a standard description of resources. We contributed to the development of several ontologies relevant to this project and have started work on the Adverse Event Reporting Ontology (AERO), presenting a prototype based on the Brighton case definition of seizure. Using queries against the created ontological model, the system can be used to guide the physician at the time of data entry, by

making sure that (i) the event they report indeed matches the Brighton case definition; (ii) we store additional information that they may not have reported upon otherwise; and (iii) we provide support to establish the correct diagnosis based on reported symptoms. Unambiguous and complete representation of adverse events following immunization will ultimately increase accuracy and quality of reporting within the PCIRN, paving the way for further adoption by Health Canada.

Availability

The latest version of the AERO file is available at: <http://purl.obolibrary.org/obo/aero.owl>.

Project home and documentation are at: <http://purl.obolibrary.org/obo/aero>.

Acknowledgments

The authors' work was partially supported by funding from the Public Health Agency of Canada / Canadian Institutes of Health Research Influenza Research Network (PCIRN), and the Michael Smith Foundation for Health Research.

OntoOrpha: An Ontology to Support the Editing and Audit of Knowledge of Rare Diseases in ORPHANET

Ferdinand Dhombres^{1,2,3,4}, Pierre-Yves Vandenbussche^{1,5}, Ana Rath², Marc Hanauer²,
Annie Olry², Bruno Urbero^{2,6}, Rémy Choquet^{1,2}, Jean Charlet^{1,2,3,7}

¹INSERM UMRS 872 éq.20, Paris, France

²INSERM SC11, ORPHANET, Paris, France

³Sorbonne Universités, UPMC, Paris, France

⁴Service de Gynécologie-Obstétrique et Centre de Diagnostic Prénatal de l'Est Parisien, Hôpital Armand Trousseau, AP-HP, Paris, France

⁵Mondeca, Paris, France

⁶INSERM DSI – Languedoc-Roussillon, Montpellier, France

⁷AP-HP – Assistance Publique, Hôpitaux de Paris, Paris, France

Abstract. ORPHANET is the reference information portal on rare diseases and orphan drugs for healthcare professionals and for general audience. After ten years of evolution, current ORPHANET tools cannot support efficiently the edition, update and data sharing processes demanded by a constantly growing rare diseases knowledge. In order to improve the editing workflow, we are conducting research to build and use a rare diseases knowledge base in an *ontology-based architecture* that complies with the W3C standards of the semantic web: OWL, RDF, SPARQL and SKOS. Our ontology design approach is based on both domain expertise (in rare diseases and in knowledge engineering) and knowledge extraction from our relational database. The current version of OntoOrpha comprises over 11,000 classes and 190,000 annotations organized under a *Rare Diseases Core Ontology*.

In comparison with the current ORPHANET editing tools, our preliminary experiments are consistent with: (1) better visualization of the knowledge base (2) improved classification editing procedures (3) improved annotation editing procedures (4) valid semantic validation procedures.

1 Background

ORPHANET is the reference information portal on rare diseases and orphan drugs for healthcare professionals and for the general public [3]. ORPHANET is led by a large European consortium of around 40 countries, coordinated by the French INSERM team which is responsible for the infrastructure of ORPHANET, management tools, quality control, rare diseases inventory, classifications and production of the encyclopedia. After ten years of evolution, current ORPHANET tools are limited in efficiently supporting the editing, update and data sharing processes of a constantly growing rare diseases knowledge (6000 rare diseases with annotations and more than one hundred overlapping classifications).

2 Methods

In order to improve the editing workflow, we are conducting research to build and use a rare diseases knowledge base in an *Ontology-based architecture*. This architecture complies with the W3C standards of the semantic web: OWL [1], RDF [4], SPARQL [9] and SKOS [7].

Our ontology design methodology is multidisciplinary as it involves both domains of expertise: rare diseases and knowledge engineering. Our methodology also involves automatic knowledge extraction from the ORPHANET relational database [5,8]. The current version of OntoOrpha comprises over 11,000 classes and 190,000 annotations, organized under a *Rare Diseases Core Ontology* (fig. 1).

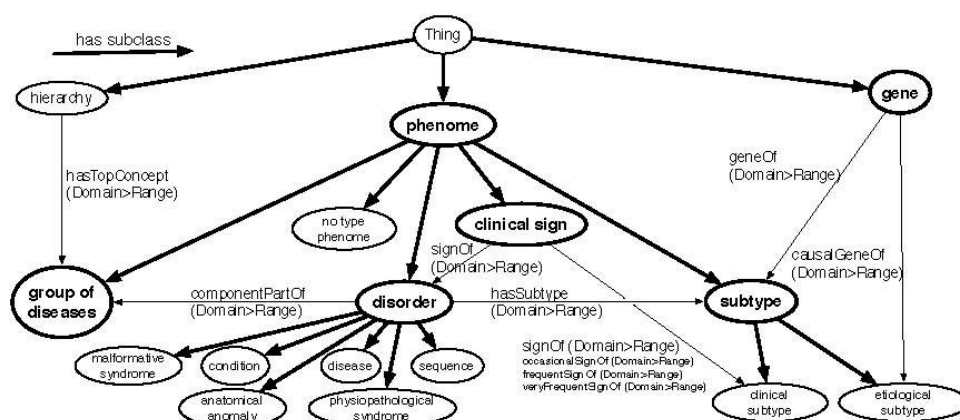


Figure1. Rare Diseases Core Ontology (view of OntoOrpha, version 2011-06-08).

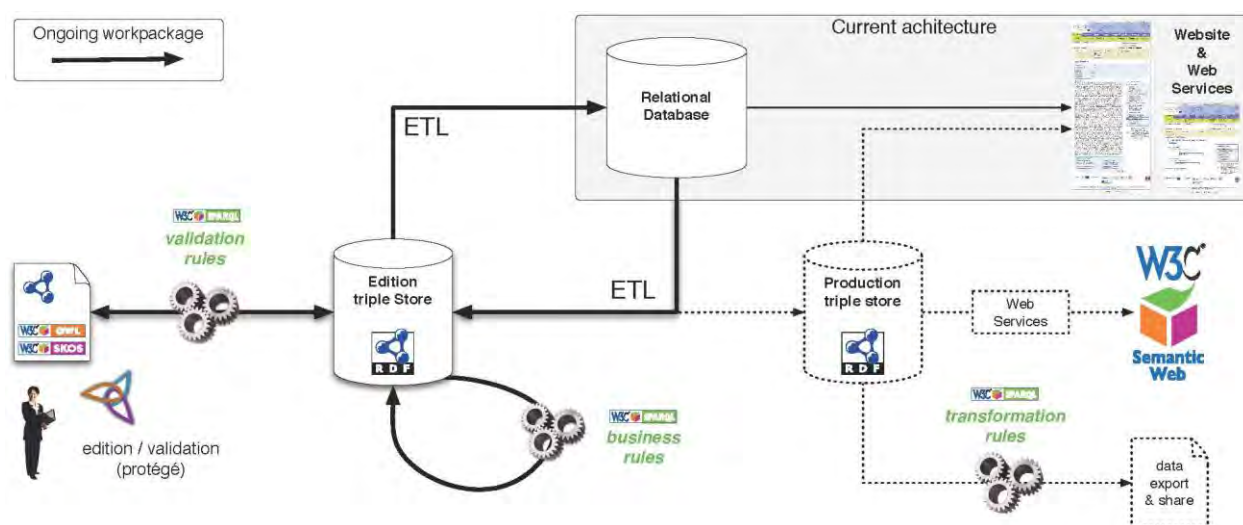


Figure 2. Rare Diseases Knowledge editing workflow (research architecture).

This core ontology was designed as a meta-model for the domain; this abstraction level was mandatory to provide the appropriate representation of the whole disease inventory extracted from the database as classes, and to represent the classifications as classes as well. Domain and range of relationships between classifications, disorders (disease, malformative syndrome, ...), groups of disorders (by anatomical system, by physiopathological mechanism, ...), subtypes (clinical subtypes, etiological subtypes), clinical signs and genes are therefore represented in the core ontology. In addition to this, we take into account that this metamodel should provide all the primitives needed for the description of validation rules.

3 Results

Besides the current architecture, a new experimental architecture is provided by the project to support editing workflow (fig. 2). At the first step of the workflow the extraction from the database, the building of the ontology and the uploading to a triplestore (semantic database) are performed. This extract-transform-load (ETL) step is automated. The generation process of the ontology is threefold:

1. generation of the header of the final RDF file (including ontology metadata, classes and object properties of the core ontology),
2. generation of the body parts by extraction/transformation of the data (URI construction, `skos:prefLabels/altLabels/`

definition including the management of 6 languages, owl:Restriction, ...)

3. merging those parts to produce an XML/RDF file that will be uploaded to the RDF triple store

The second step is editing the ontology by the expert (Classes, Object-Property, Annotations) and rule-based procedures implemented with iterative SPARQL queries on the triplestore (for validation, classification generation, audit report generation). The final step is the relational database update with an ETL process from the triplestore.

In comparison with current ORPHANET editing tools, our preliminary experiments are consistent with:

- a better visualization of the knowledge base : a global view of the hierarchies is provided in the ontology editor during the edition process (*Class hierarchy view*, *Existential tree view*, *Outline tree view* and *OntoGraf view* in PROTÉGÉ [2]),
- improved classification editing procedures: the experts edit the ontology itself and the rare diseases classifications are automatically generated, using stable SPARQL queries,
- improved annotation editing procedures: we use a lexicalization plugin for PROTÉGÉ developed in our unit [6] that is SKOS compliant and supportive for multilingual editing of labels, synonyms and abstracts,
- semantic validation procedures: both in the ontology editor (*a priori* validation by the build-in reasoner) and in the triplestore (validation and audit by reasoner and SPARQL queries).

Finally, the core ontology allows us to globally review and reorganize the ORPHANET rare disease knowledge. It provides the necessary coherent top-structure of the

knowledge managed into the knowledge base (diseases, classifications, genes, ...). ORPHANET core ontology guarantees a consistent evolution of the ontology going forward.

References

1. OWL 2 web ontology language. W3C recommendation. W3C OWL working group. (2009), <http://www.w3.org/TR/2009/REC-owl2-overview-20091027/>
2. Protégé 4 - Open Source Ontology Editor. Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine (2011), <http://protege.stanford.edu/>
3. Aymé, S.: Orphanet: serveur d'information sur les maladies rares et les médicaments orphelins, INSERM SC11. <http://www.orpha.net/> (2002).
4. Beckett, D., McBride, B.: RDF/XML syntax specification (Revised). W3C recommendation. (2004), <http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/>
5. Krivine, S., Nobécourt, J., Soualmia, L., Cerbah, F., Duclos, C.: Construction automatique d'ontologie à partir de bases de données relationnelles: application au médicament dans le domaine de la pharmacovigilance. In: Actes des 20e Journées Francophones d'Ingénierie des Connaissances. pp. 73–84. Fabien Gandon (Ed.), Hammamet, Tunisie (2009)
6. Mazuel, L.: Archonte plugin for protégé. (2010), <http://pertomed.spim.jussieu.fr/~lma/doe/fr.spim.archonte.jar>
7. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. W3C recommendation. (2009), <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>
8. O'Connor, M.J., Das, A.: Semantic reasoning with XML-based biomedical information models. *Studies in Health Technology and Informatics* 160(Pt 2), 986–990 (2010)
9. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF. W3C recommendation. (2008), <http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/>

An Ontology for Gastrointestinal Endoscopy

Shahim Essaid

Oregon Health & Science University, Portland, OR, USA

Abstract. The field of gastrointestinal endoscopy can benefit from an ontology for the purposes of data coding and data integration. This paper presents early results of an effort to develop such an ontology based on the OBO Foundry principles and existing OBO Foundry ontologies. Initially, the ontology will be limited to representing entities and relations currently implicit or hard-coded in the user interface of an existing endoscopy reporting system. The ontology will also be mapped to an existing database to evaluate the feasibility of ontology-driven queries. The long-term goal is to evolve this ontology to an application-independent terminology and information model for the domain of gastrointestinal endoscopy.

Keywords: gastrointestinal endoscopy, knowledge representation, data integration

1 Introduction

The practice of gastrointestinal endoscopy produces a significant amount of structured data that is captured in endoscopy reports. To encourage consistency in data collection, the World Endoscopy Organization maintains the Minimal Standard Terminology for gastrointestinal endoscopy (MST) [1]. The MST specifies a minimal set of terms and data structures needed to encode the majority of endoscopy data. However, the domain knowledge represented in the MST (in the form of terms, relations, and data structures) is not in a computable format and it could benefit from an ontological and logical analysis, and reorganization.

Another important effort in the field of gastrointestinal endoscopy is the Clinical Outcomes Research Initiative (CORI) [2]. CORI was established to assess the utilization and effectiveness of endoscopy procedures in clinical care. To meet its goals, CORI has developed an endoscopy reporting software and a central data repository for endoscopy reports. The reporting software is being used nationwide and the data repository currently receives over 250,000 reports annually. The data repository is primarily used for research purposes and to report on practice patterns and clinical outcome measures.

The endoscopy reporting software developed by CORI was initially based on the

MST's representation of endoscopy data but it evolved to include additional terminology and data elements. Also, efforts are in place to integrate data generated by commercial reporting systems into the CORI data repository. These efforts have highlighted the need for a shared, stable, and computable terminology and information model for the field of gastrointestinal endoscopy to facilitate data integration while maintaining clear and consistent semantics.

2 Motivation and Planned Development

Recent advances in the area of biomedical ontology [3-4] can provide a foundation for a more formal and logical representation of entities and data elements needed to represent endoscopy data. Also, the existence of standardized knowledge representation languages and related inferencing capabilities can enable sophisticated querying of logically represented data and knowledge [5]. These advances have motivated an effort within the CORI project to develop an ontology for the field of gastrointestinal endoscopy.

The ontology will follow the Open Biomedical Ontology (OBO) Foundry development principles [6] and reuse entities from existing OBO ontologies when appropriate. The BFO will serve as an upper level ontology and other ontologies (IAO,

ogms, OBI, etc.) will be examined for middle level entities. Domain level entities will reference existing ontologies of anatomy, pathology, phenotypes, relations, and others when available. The ontology development project is hosted as a Google Code project [7].

3 Methods and Expected Difficulties

Development will start by identifying domain level terms and data elements hard-coded in the user interface of the existing CORI reporting software. These entities will initially form the main content of the ontology. This will decouple the domain knowledge from the application and allow for a more flexible evolution of the terminology and information model of the reporting software while still maintaining ontological and formal knowledge representation principles. The ontology will then be augmented by other entities from the MST, from free text entries in existing endoscopy reports, and from the endoscopy community. Also, as a proof of concept, the ontology will be mapped to the existing CORI data repository to evaluate the feasibility and benefit of ontology-driven queries compared to native SQL queries. The D2RQ Platform will be used for this part of the project [8].

A brief exploration of the user interface for the reporting software, and the MST, showed that difficult issues such as epistemology vs. ontology, entities vs. statements, negation, and other related issues are frequent in clinical settings. Also, despite the relatively narrow focus of the practice of gastrointestinal endoscopy, endoscopy reports include information that ranges from current and past medical history, physical examination, visible endoscopy findings, and indirect findings through various imaging techniques. In addition to these various types of information, there is another epistemic layer that reflects the attitudes and judgment of clinicians in the form of assessments, diagnoses, plans, etc. To fully represent this information, an ontology will need a rich set of relationships that cover mereotopological, temporal, and modal relations, among others.

However, the primary use cases described below can be met by limiting our initial development efforts to an *is_a* hierarchy and a basic set of qualitative mereotopological relationships. The initial version will also be limited to representing endoscopic findings (polyps, ulcers, foreign bodies, etc.), their anatomical locations, and their clinical descriptions, according to the OBO Foundry ontology development principles.

4 Primary Use Case

The ontology will primarily serve as an interface terminology that supports data entry and enables consistent coding of endoscopy reports. The ontology will also be used to explore the value of ontology driven data retrieval by executing ontology driven queries against the current CORI dataset. These initial use cases can be met by a limited set of entities and relations and the remaining domain knowledge will be added as need arises.

References

1. Minimal Standard Terminology, <http://www.worldendo.org/mst.html>
2. Clinical Outcome Research Initiative, <http://www.cori.org/>
3. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform.* 7(3):256–274 (2006)
4. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25(11):1251–1255 (2007)
5. Stevens R, Aranguren ME, Wolstencroft K, Sattler U, Drummond N, Horridge M, et al. Using OWL to model biological knowledge. *International Journal of Human-Computer Studies.* 65(7):583–594 (2007)
6. OBO Foundry Principles, <http://www.obofoundry.org/crit.shtml>
7. Gastrointestinal Endoscopy Ontology, <http://giao.googlecode.com/>
8. The D2RQ Platform – Treating Non-RDF Databases as Virtual RDF Graphs, <http://www4.wiwi.fu-berlin.de/bizer/d2rq/>

Enriching the Ontology for Biomedical Investigations (OBI) to Improve its Suitability for Web Service Annotations

Chaitanya Guttula¹, Alok Dhamanaskar¹, Rui Wang¹, John A. Miller^{1,3},
Jessica C. Kissinger^{1,2,3}, Jie Zheng⁴, Christian J. Stoeckert, Jr.⁴

¹Department of Computer Science, ²Department of Genetics, ³Institute of Bioinformatics,
University of Georgia, Athens, GA, USA

⁴Penn Center for Bioinformatics, Department of Genetics University of Pennsylvania, Philadelphia, PA, USA

Abstract. With the increasing development and use of ontologies in the biomedical domain, opportunities for their utilization in applications and workflows are being created. In this paper, we discuss how the Ontology for Biomedical Investigations (OBI) can be enriched to support annotation of Web services. The methodology includes designing ontology analysis diagrams for Web services and analyzing them to find the terms that need to be added to the ontology. The enriched ontology can then be used for annotating the Web services with the help of annotation tools like the one in the RadiantWeb tool-suite. Using annotated Web services to perform service discovery and make service suggestions provides a way to evaluate the validity of the annotations made and the terms added.

Keywords: ontology, OBI, biomedical, Web services, semantic annotations.

1 Introduction

In recent years, the number of tools and software applications available as Web services in the biomedical community has increased dramatically. Complex real world tasks generally require coordinated use of multiple Web services. It is a challenge to find those Web services that suit the users' needs or work effectively in Web service compositions. Semantic annotations of the Web services would facilitate Web service discovery and composition [1]. A Web service may be described using the Web Service Description Language (WSDL), which specifies the set of operations provided by the Web service, as well as details about these operations, including their inputs and outputs. Standardized annotations of a Web service include the semantics for the input, output and functionality of each of the service's operations. Bioinformatics Web services are used to analyze biomedical data and hence, need relevant terms for their annotation.

It is preferable to use an ontology that is compatible with other biomedical ontologies. Open Biological and Biomedical Ontologies [2] (OBO) compliant ontologies are interoperable with each other, because they share a common

upper level ontology, the Basic Formal Ontology (BFO) [3], and a common set of relations, the Relation Ontology (RO) [4]. The Ontology for Biomedical Investigations (OBI) [5], a member of the OBO Library, is being developed to address the need for consistent description of biological and clinical investigations. OBI is a process oriented ontology that models a process with input, output and objective specifications and is suitable for supporting Web service annotations. This paper reports on our efforts to enrich OBI for the purpose of semantically annotating Web services to enhance its usability.

2 Enrichment of OBI to Support Service Annotations

Ontologies used for annotations should provide terms that correspond to key aspects of a Web service description. If the required terms are not available in OBI, we add them. However, terms are reused where possible. We begin the process of enriching OBI by creating a generic model for Web services and further refine it to model specific types of Web services. Constructing a generic model involves creating an ontology analysis diagram, which shows

relationships between different top level terms for Web services, including the objective of a Web service and its operations.

A sample generic model can be viewed at: <http://mango.ctegd.uga.edu/jkissingLab/SWS/Wsannotation/resources/GenericCmap.jpg>

Modeling a Web service at a more specific level requires a detailed analysis of the Web service's operations in terms of its inputs, outputs and objective specification. The outcome of this can be seen for example in the ontology analysis diagram for ClustalW, available at: <http://mango.ctegd.uga.edu/jkissingLab/SWS/Wsannotation/resources/clustalCmap.jpg>.

We are modeling several Web services, including ClustalW and BLAST Web services currently available at the European Bioinformatics Institute (EBI). The ClustalW Web service was studied and its inputs and outputs were summarized in a spreadsheet along with

their definitions. We finally determine the terms and the possible positions where they can be added to the ontology on the basis of the above-mentioned ontology analysis diagram. Once a term is fully described by specifying its set of restrictions (e.g., objective specification for Web service operations), we come up with a logical definition for the term.

Using Protégé, we have added the new terms in the OWL file that is available at: obi.svn.sourceforge.net/svnroot/obi/trunk/src/ontology/branches/webService.owl (Figure 1). A description logic reasoner (e.g., HermiT) is used to check for consistency of the added terms, as well as to infer the correct placements of the terms in the ontology's hierarchy. A request has been sent to the OBI issue tracker and the terms are currently pending approval.

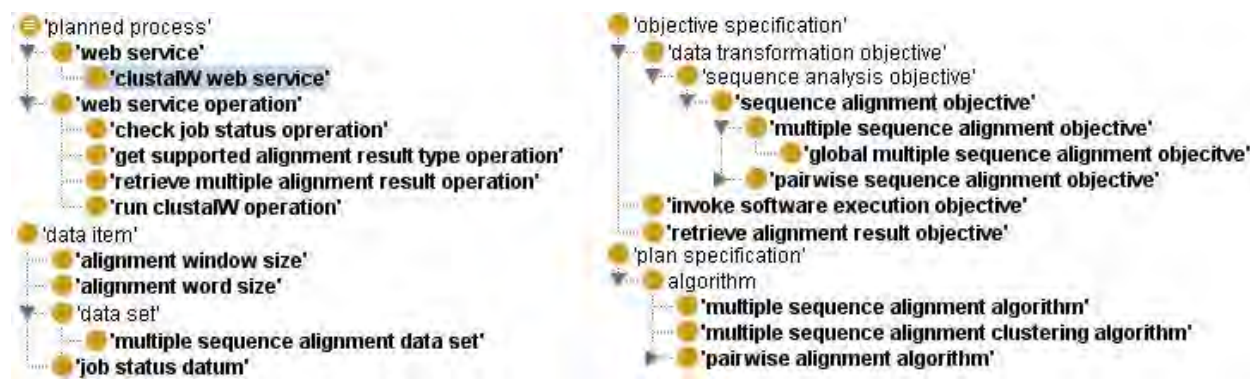


Figure 1. Ontology Hierarchy of terms added to OBI used for clustalW annotation, the terms in bold are the newly added terms

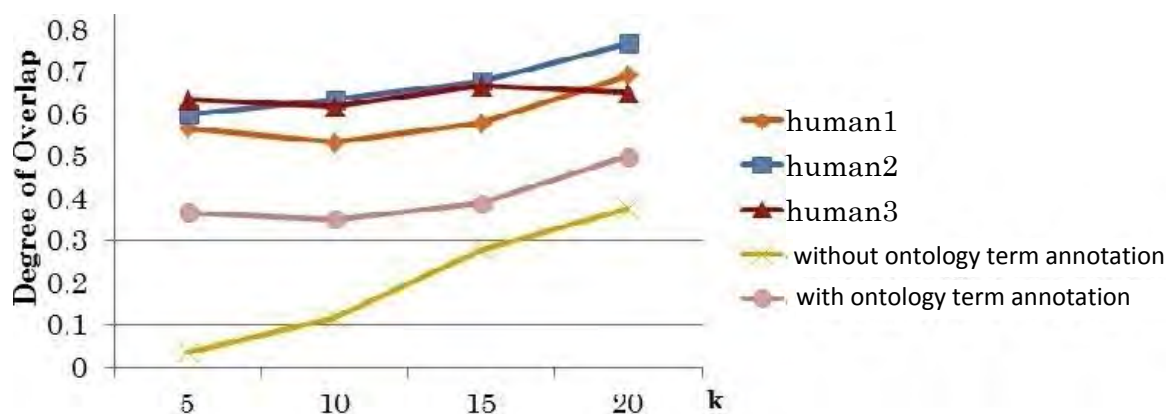


Figure 2. Comparing Annotation Cases.

3 Evaluations of Web Service Annotations Using OBI

Considering the increasing number of available Web services in Biomedical domain, manual annotation and composition of Web services is a tedious task. The RadiantWeb tool-suite is a Web application with a simple drag-and-drop user interface that includes tools for annotation, discovery and suggestion of Web services. The purpose of annotations is to provide formalized documentation that can be read by humans and processed by machines.

To illustrate the use of annotations, we considered a common scenario encountered by biologists, that of discovering more information about a particular protein sequence and its evolutionary relationship to other protein sequences. In order to find this information, we had to design a workflow consisting of multiple Web service operations. The workflow mainly utilized popular bioinformatics programs such as BLAST and ClustalW. The RadiantWeb Tool Suite was used to ease the process of providing annotations and creating the workflow. Figure 2 depicts the effectiveness of annotations, which confirms that the annotated Web services perform better for service discovery and suggestions than unannotated ones. A more detailed evaluation of effectiveness of annotations can be found in [7].

4 Discussion: Related Work & Conclusions

To the best of our knowledge the only other major effort in the biomedical domain focused on creating or enriching ontologies for the purpose of semantically annotating Web services is the EMBRACE Data and Methods (EDAM) ontology [6]. EDAM covers several but not all of the terms required for the annotation of Web services in this domain. For example, for BLAST Web Services missing terms include ‘low complexity sequence filter’, ‘number of top combinations’, ‘pairwise alignment sensitivity’. Also, EDAM is a work in progress with several of its properties having ranges/restrictions specified as undefined, in addition to it not being OBO compliant,

meaning it has lesser compatibility with other biomedical ontologies that are OBO compliant.

In this paper, we apply a systematic methodology for enriching existing biomedical ontologies (OBI in our case), so that they can support semantic annotation of Web services in this domain. We have enriched OBI with terms required for the annotation of BLAST & ClustalW and will continue working on other Web services. Tools such as RadiantWeb can be used to make the process of annotating Web services quick and easy. Our preliminary evaluation made it clear that discovery and service suggestions with annotated Web services yield better results than with unannotated Web services [7].

Acknowledgements

Funding for this study was provided by NIH R01 GM093132.

References

1. Phillip Lord, Sean Bechhofer, et al.: Applying Semantic Web Services to Bioinformatics. *LCNS*, vol. 3298, pp. 350-364 (2004).
2. Barry Smith, Michael Ashburner, et al.: The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature Biotechnology* vol. 25, pp. 1251-1255 (2007).
3. <http://www.ifomis.org/bfo>
Accessed Apr 30, 2011.
4. <http://www.obofoundry.org/ro/>
Accessed Apr 30, 2011.
5. Brinkman RR, Courtot M. et al.: Modeling Biomedical Experimental Processes with OBI. *Journal of Biomedical Semantics*, Suppl 1:S7 (Jun 2010).
6. <http://edamontology.sourceforge.net/>
Accessed Apr 30, 2011.
7. Rui Wang, Chaitanya Guttula, Maryam Panahiazar, Haseeb Yousaf, John A. Miller, Eileen T. Kraemer and Jessica C. Kissinger: Web Service Composition using Service Suggestions. In: *Proceedings of the 2011 IEEE International Workshop on Formal Methods in Services and Cloud Computing*, Washington, DC (July 2011).

Recent Developments in the ChEBI Ontology

Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Kenneth Haug,
Zara Josephs, Gareth Owen, Steve Turner, Christoph Steinbeck

Chemoinformatics and Metabolism, European Bioinformatics Institute, Cambridge, UK

Introduction

ChEBI – Chemical Entities of Biological Interest – is an ontology of chemical entities such as molecules and ions, and their roles in biological contexts [1]. As of April 2011, it contains in total around 25,000 classes. Here, we report on recent developments and changes in the ontology, and give a brief view on ongoing work that will lead to changes in the future.

1 Recent changes

1.1 Mapping to upper-level ontology BFO

In order to comply with our goal of increasing interoperability with other ontologies in the biomedical domain, ChEBI has undertaken to provide a mapping to the upper level ontology BFO (Basic Formal Ontology) [2]. Mapping multiple ontologies beneath a common upper level allows easier linking between ontologies, since it reduces ambiguities in interpretations through a clear ontological commitment.

The ChEBI mapping to BFO is illustrated in Figure 1, and provided as an OWL file, which is downloadable from:
<ftp://ftp.ebi.ac.uk/pub/databases/chebi/ontology/>

1.2 Renaming of ‘Molecular Structure’ Root

ChEBI renamed the root term of the sub-ontology in which chemical entities such as molecules and ions are defined, from ‘molecular structure’ to ‘chemical entity’. While the historical name accurately reflected the organising principle of the sub-ontology (the classification of entities therein is on the basis of structural features), it was not adequate for purposes of automated reasoning, since it led to incorrect inferences through the transitivity of the *is a* relationship,

such as:

caffeine *is a* molecular structure.

After the modification, we have the correct inference:

caffeine *is a* chemical entity.

1.3 Expanded Substance Hierarchy

In order to adequately deal with user-requested mixtures and polymers within the ontology, ChEBI has expanded its ‘chemical substance’ hierarchy. This reflects a slight change in scope relative to earlier versions of the ontology, which tried to explicitly exclude aggregate chemical substances. The changes has been introduced to allow adequate classification of some of the entities which were strongly requested by our users, and for which no other suitable ontology yet existed.

We have created a new upper-level term beneath ‘chemical entity’: *chemical substance*. We further differentiate between pure and mixed substances. An example of a pure substance is a macroscopic homogeneous collection of molecular entities (such as, say, water), while a mixture contains a non-homogeneous collection – composed of at least two different sorts of entity. In particular, this allows us to correctly model racemic mixtures, which are crucial in the adequate representation of drugs, since many active substances found in drugs are formulated as racemic mixtures. Most chemical databases skirt the issue of representing racemic mixtures, or do so inconsistently. (For a discussion of this point, see: <http://chem-bla-ics.blogspot.com/2011/02/chemical-data-curation-yes-it-is-that.html>)

Our preferred ontology representation for racemic mixtures is illustrated in Figure 2.

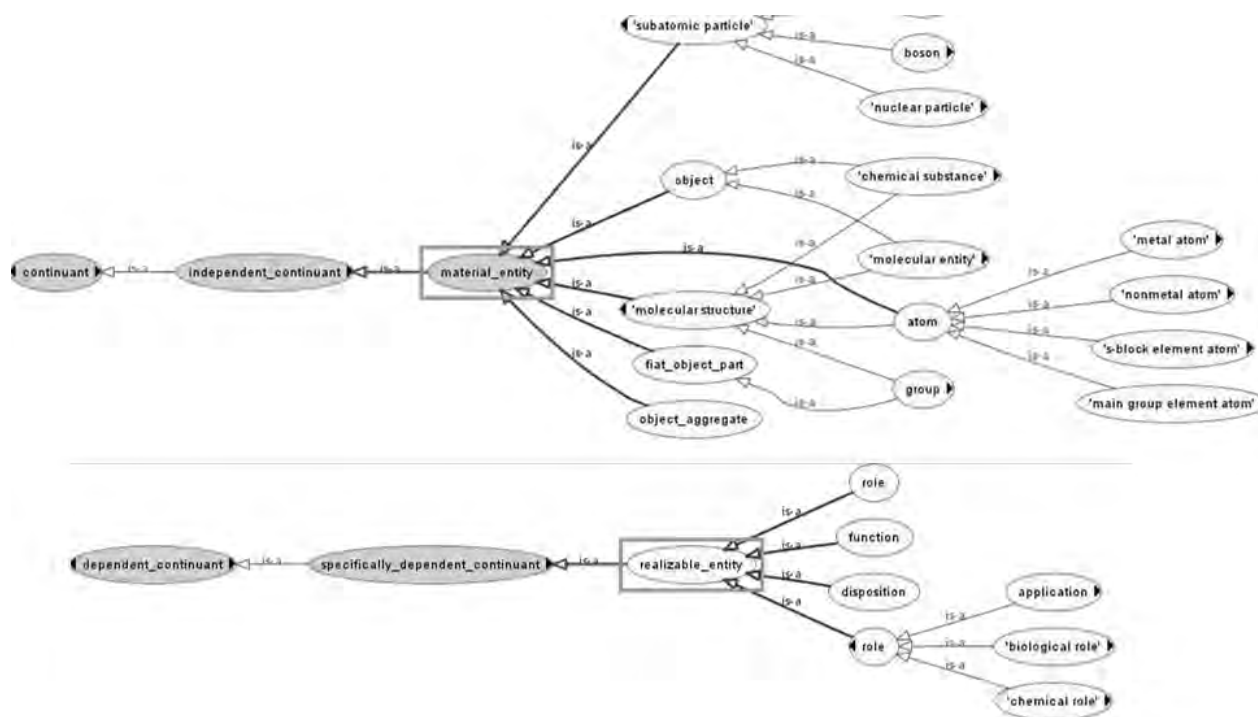


Figure 1. BFO Mapping.

The figure illustrates the mapping between upper-level ChEBI classes and the BFO terms to which they map, separated between ChEBI chemical entities and subatomic particles as independent continuants and ChEBI roles as realizable entities. Note that ChEBI 'role' does not map to BFO:role but to BFO:realizable entity.

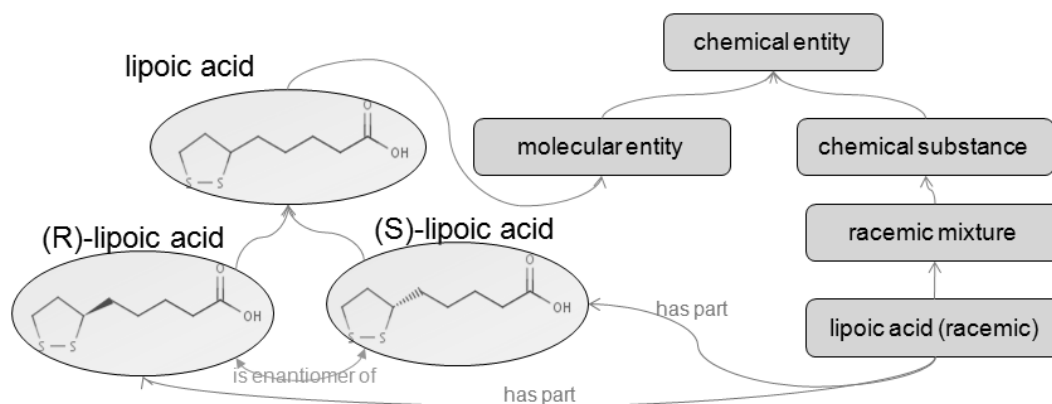


Figure 2. Racemic mixtures: Mixtures in ChEBI are explicitly modelled and their parts are linked via 'has part' relations.

1.4 Refactoring Natural Products

Natural products are of substantial interest in drug discovery and metabolism research, since they represent molecules that in many cases have been shaped by natural selection to be bioactive in highly specific ways. However, the core definition of what specifically constitutes a natural product is seldom rendered explicit, and differs from community to community. Some

candidate definitions among the many possibilities, in sequence from more inclusive to more exclusive, are:

1. All chemicals that can be isolated from a living organism;
2. All metabolites (primary and secondary);
3. Secondary metabolites only;

4. Secondary metabolites in plants only.

ChEBI currently includes classes related to natural products in two different places in the ontology. Firstly, common natural product families are explicitly classified in the chemical entity ontology, and secondly, ‘metabolite’ is specified in the role ontology. An example of a natural product class in the chemical entity ontology is:

cinchonine *is a* heterocyclic natural product
is a natural product.

The text definition¹ is as follows: “Cinchonan in which a hydrogen at position 9 is substituted by hydroxy (S configuration). It occurs in the bark of most varieties of Cinchona shrubs, and is frequently used for directing chirality in asymmetric synthesis.”

Due to the inherent ambiguity, current curation efforts involve the deprecation of classes explicitly containing ‘natural product’ in their name, and instead classifying molecules such as cinchonine explicitly as secondary metabolites. Future work will involve adding the species as explicit context to the definition of metabolites.

2 Ongoing Work and Future Changes

2.1 Focus on immunology

A large-scale ongoing curation effort in collaboration with the La Jolla Institute for Allergy and Immunology (LIAI, <http://www.liai.org/>) is focused on annotating compounds relevant for immunology, such as those which act as antigens and immunogens. ChEBI has so far annotated more than 1,000 such compounds.

2.2 Relationship Definition and Re-Evaluation

ChEBI is undergoing a major re-evaluation of the relationships which it makes use of, in order to bring them in line with the RO [3] where possible, and to provide formal definitions for chemistry-specific relationships. As part of this ongoing process, ChEBI will introduce RO relations such as *disjoint from*, and may

deprecate some of the chemistry-specific relations such as *has parent hydride* if they prove resistant to full logical definition.

2.3 Disentangling Role and Chemical Entity

Prior to 2009, the *is a* relationship in ChEBI was overloaded, linking molecular entities with chemical classes and specifying the ‘roles’ that chemical entities can enact in various contexts. To address this, the relationship *has role* was introduced and used to link molecular entities to roles, for example, the molecular entity acetylsalicylic acid (CHEBI:15365) *has role* non-narcotic analgesic (CHEBI:35481). The initial disentanglement was performed programmatically, and subsequent manual curation was required to clean up some cases where errors occurred, such as when a chemical entity lacked a structure and was only classified with a role parent. Current curation efforts are underway to fully define classes which are specified with both structural and role-based features, such as the entity tricyclic antidepressant (CHEBI:36809), which is defined as *is a* organic tricyclic compound and *has role* antidepressant.

References

1. de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C.: Chemical Entities of Biological Interest: an update. *Nucl. Acids Res.* 38, D249–D254 (2010)
2. Grenon, P., Smith, B.: SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition & Computation: An Interdisciplinary Journal* 4(1), 69–104 (2004).
3. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., Rosse, C.: Relations in biomedical ontologies. *Genome Biology* 6(R46) (2005)

¹ Sourced from ChEBI version 78, <http://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:27509>

Representing Local Identifiers in a Referent-Tracking System

William R. Hogan¹, Swetha Garimalla¹, Shariq A. Tariq¹, Werner Ceusters²

¹Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

²Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY, USA

Keywords: realist ontology, referent tracking, electronic health records

1 Introduction

Referent Tracking (RT) is a novel, principled approach, suitable to a diversity of applications, to store data about particulars in reality. RT is unique in that it (1) assigns globally unique singular identifiers for *each* entity in reality about which information is stored, rather than for only some obvious entities such as, in the case of electronic health records (EHRs), the patient and his caregivers, and (2) uses a series of templates for unambiguous representation of the relationships of particulars [1-4].

At the University of Arkansas for Medical Sciences (UAMS), we are investigating RT's ability to handle diagnoses, procedures, demographics, encounters, hypersensitivity, and observations as they are reported in EHRs. The questions we are addressing are, amongst others, ones of representational adequacy: is it possible to represent in a referent tracking system (RTS) – with the current facilities the RT approach provides – the same entities in reality that EHR data are about?

A problem at this time, however, is that there are no such EHR systems. Existing EHRs – like numerous other biomedical software applications – use unique internal identifiers¹ to denote only those particulars that are persons, healthcare encounters, medical records, etc., but not diseases, injuries, disease courses, tumors, etc. For the sake of interoperability with non-RTS EHRs that cannot store mappings between local identifiers and IUIs, we require the ability to include local identifiers in the RTS. In this work, we

confronted the issue of how to handle local identifiers in an RTS without violating the underlying ontological principles. The essence of the approach is to represent local identifiers just as we represent other entities external to the RTS: with instance unique identifiers (IUIs). Doing so however raised additional issues. The question we address here is whether we can solve these issues using existing RTS capabilities as embedded in the RT templates.

We use the following hypothetical scenario to illustrate our approach by building up a set of RT templates that represents the scenario: *Mrs. Smith is a new patient at ABC Medical Clinic, where an EHR has been in operation since 2005-05-05. On 2011-01-01, she checks in at the front desk and a member of the clinic staff enters her basic demographic and insurance information and creates a medical record for her in the EHR. After her visit with Dr. Jones, Mrs. Smith checks out and leaves the clinic.*

In our scenario, we assume (1) appropriate distinction between person identifiers and medical record numbers (MRNs), (2) the use of a single person identifier regardless of participation in an encounter as doctor or patient, (3) that the identifier 'EHR00001'² refers to the EHR instance at ABC Medical Clinic and is a local identifier in that EHR, and (4) that the local identifier 'O00001' uniquely denotes the organization. We ignore the building in which ABC Medical Clinic operates as it is not essential to our scenario, although this entity too requires a unique local identifier.

¹ We will from now on use the term 'local identifier' for the identifiers used in EHR systems to distinguish them clearly from the identifiers used in a referent tracking system.

² A note on use vs. mention: when we mention an identifier, we set it in single quotes. On the other hand, when we use an identifier to refer to something, we set it in italics.

2 The Approach

The approach is based on the fact that local identifiers and systems of such identifiers as used in healthcare organizations are as real and as external to the RTS as are the entities that they denote and can thus be denoted by IUIs too. In what follows, we use shorthand notation for IUIs, such as IUI_{Smith} , to improve readability.

We first assign IUIs to entities in the scenario using the RTS assignment or A-template: Mrs. Smith, Dr. Jones, ABC Medical Clinic, the EHR, and Mrs. Smith's medical record: $A < IUI_{Smith}, IUI_a, t_{ap} >$, $A < IUI_{Jones}, IUI_a, t_{ap} >$, $A < IUI_{ABC}, IUI_a, t_{ap} >$, $A < IUI_{EHR}, IUI_a, t_{ap} >$ $A < IUI_{SmithRecord}, IUI_a, t_{ap} >$. The identifiers IUI_a and t_{ap} denote the entity that made the IUI assignment and the time of the assignment, respectively. See Ceusters [1] for a complete description of RT templates; here we provide enough explanation to illustrate our approach.

Each local identifier belongs to some system of identifiers. The system of identifiers is almost always constructed with the goal that each identifier in the system has one, unique, unambiguous reference. In our scenario, we included five identifier systems, because in our experience most EHRs have different systems for persons, encounters, etc. However, one system is feasible so long as all its identifiers uniquely denote one entity (for example, no encounter identifier would have the same string of characters as an MRN). What follows is the representation of the person identifier system in the EHR. We treat each of the other four systems (encounter, MRN, organization, and EHR identifier) in the same manner.

First, we assign an IUI to the person identifier system: $A < IUI_{PersonIdSystem}, IUI_a, t_{ap} >$ and assert that it is an instance of a central identifier registry using a Particular-to-Universal (PtoU) template and the appropriate universal representation from the Information Artifact Ontology (IAO): $PtoU < IUI_a, t_a, inst, http://purl.obolibrary.org/iao, IUI_{PersonIdSystem}, IAO_0000579, t_r >$. So that humans can differentiate among the various identifier systems represented in the RTS, we assign a name to each one using a Particular-to-Name (PtoN) template, which has the form: $PtoN < IUI_a, t_a, IUI_c, IUI_p, n, nt, t_r >$ where IUI_c

is the IUI for the entity that uses the name n , IUI_p is the IUI for the entity with the name n , nt is the name type (e.g., first name), and n the name associated with IUI_p . For the person identifier system, we use $PtoN < IUI_a, t_a, IUI_{ABC}, IUI_{PersonIdSystem}, 'internal system name', 'Person id system', t_r >$. We similarly assign each local identifier an IUI, assert that it is an instance of centrally registered identifier, and assign it a name (the identifier string is the name n , e.g. '000001'). Each local identifier is part of its identifier system and denotes some entity, and each identifier system is part of the EHR: we represent these relationships for each identifier and system using Particular-to-Particular (PtoP) templates. The full set of templates for the scenario is publicly available online as a Google document: <https://spreadsheets.google.com/ccc?hl=en&key=t8v6oS7tNl84OMDjuy5p2kQ&hl=en#gid=0>

3 Discussion

We successfully represented EHR local identifiers in an RTS using existing RT facilities. Our approach is general, and could be used to represent local identifiers and identifier systems in any non-RTS.

The approach has certain advantages. Besides the already mentioned disambiguation of what exactly is denoted by a local identifier, distinct units within one organization can continue to use local identifiers despite referencing the same entities, and this without the need for complex identity-negotiation systems [7], or the need for an a priori agreement on a fixed set of entity types [8]. Thus it solves many issues the traditional federated database approach suffers from [9]. When EHRs of distinct organizations that provide healthcare to overlapping patient populations are connected to the same RTS or to RTSs which are connected in an RTS network [3], the approach enables tracking of the variety of identifiers used within these organizations. And when extended to include local dictionaries within units or organizations, the approach provides the additional benefit of implementing Smith's proposal to counteract the drawbacks of traditional controlled vocabularies and terminologies by using EHR data as a means to quality-control them (and

thus for purposes of automatically generating improved versions of such dictionaries) [10].

A drawback of the approach is that the PtoN-template, and more specifically the “name type” slot of that template, might become overloaded in its own right, and that at some point a name-type system might become necessary to track the various sorts of name types in use. Also, the approach leaves a number of relationships implicit, for example, that the systems of identifiers are endorsed by the organizations in whose EHRs local identifiers thereof are used. This problem could, in a naïve way, be solved by adding additional PtoP templates for which rather ad hoc relationships such as ‘endorses’ need to be defined. This sort of solution clashes however with the principles of Ontological Realism [11] to which RT aims to adhere. A better approach, and the topic of future work, is to introduce denotational bonds as proposed by Ceusters [12].

4 Conclusion

We identified a need to represent local identifiers and systems of such identifiers in EHRs in our work on RT. Prior to this work, whether and how RT could enable such a representation was an open question. The answer was affirmative: we successfully developed the required representations in an RTS and that made use of existing RT facilities. The approach nevertheless has some limitations we intend to address in future work by developing an ontological theory of denotational bonds.

References

1. Ceusters W. Dealing with mistakes in a referent tracking system. *Ontology for the Intelligence Community (OIC-2007)*; Columbia, Maryland. 2007. p. 5-8.
2. Ceusters W, Elkin P, Smith B. Negative findings in electronic health records and biomedical ontologies: a realist approach. *Int J Med Inform.* 2007;76 Suppl 3:S326-33.
3. Ceusters W, Manzoor S. How to track absolutely everything. In: Obrst L, Janssen T, Ceusters W, editors. *Ontologies and Semantic Technologies for Intelligence*. Amsterdam: IOS Press; 2009.
4. Ceusters W, Smith B. Strategies for referent tracking in electronic health records. *J Biomed Inform.* 2006 Jun;39(3):362-78.
5. Rudnicki R, Ceusters W, Manzoor S, Smith B. What particulars are referred to in Electronic Health Record data? A case study in integrating Referent Tracking into an EHR application. *AMIA Annu Symp Proc.* 2007:630-4.
6. Ceusters W, Smith B. Referent tracking for digital rights management. *International Journal of Metadata, Semantics and Ontologies.* 2007;2(1):45 - 53.
7. Meghini C, Doerr M, Spyrtos N. Managing co-reference knowledge for data integration. In: Kiyoki Y, Tokuda T, Jaakkola H, Chen X, Yoshida N, editors. *Information Modelling and Knowledge Bases*. Amsterdam, The Netherlands: IOS Press; 2009. p. 224-44.
8. Bouquet P, Stoermer H, Niederee C, Maa A. Entity Name System: The back-bone of an open and scalable web of data. *Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008*, number CSS-ICSC 2008-4-28-25: IEEE Computer Society; 2008. p. 554-61.
9. Heimbigner D, McLeod D. A federated architecture for information management. *ACM Trans Inf Syst.* 1985;3(3):253-78.
10. Smith B, Ceusters W. An ontology-based methodology for the migration of biomedical terminologies to electronic health records. *AMIA Annu Symp Proc.* 2005:704-8.
11. Smith B, Ceusters W. Ontological realism as a methodology for coordinated evolution of scientific ontologies. *Applied Ontology.* 2010;5(3-4):139-88.
12. Ceusters W. Introduction to Referent Tracking. Tutorial co-located with the International Conference on Biomedical Ontology (ICBO), Buffalo NY and part of the 2-day class From Basic Formal Ontology to the Information Artifact Ontology, 2009: Available from: http://www.bioontology.org/wiki/index.php/Ontology_Training:_Video_and_Audio_Presentations.

owl_cpp, a C++ Library for Working with OWL Ontologies

Mikhail K. Levin¹, Alan Ruttenberg^{2,3}, Anna Maria Masci⁴, Lindsay G. Cowell¹

¹Department of Clinical Sciences, University of Texas Southwestern Medical Center at Dallas, TX, USA

²School of Dental Medicine, University at Buffalo, NY, USA

³Science Commons, Mountain View, CA, USA

⁴Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA

Abstract. Here we present *owl_cpp* (<http://sf.net/projects/owl-cpp/>), an open-source C++ library for parsing, querying, and reasoning with OWL 2 ontologies. *owl_cpp* uses Raptor, FaCT++, and Boost libraries. It is written in standard C++, and therefore can be used on most platforms. *owl_cpp* performs strict parsing and detects errors ignored by other parsers. Other advantages of the library are high performance and a compact in-memory representation.

1 Introduction

The OWL Web Ontology Language, one of the Semantic Web technologies, is designed to formally represent human knowledge and facilitate its computational analysis and interpretation. OWL and other Semantic Web technologies have been successfully applied in many areas of biomedicine. Their success is in part due to a broad spectrum of software developed in support of the Semantic Web.

Working with ontologies involves several common tasks, such as parsing OWL documents, querying their in-memory representation, passing information to a Description Logic (DL) reasoner, and executing DL queries. In computer memory ontologies are represented either as RDF triples or as axioms and annotations. A triple is a simple statement consisting of three nodes, where the predicate node expresses a relationship between the subject and object nodes. Although a set of triples can represent a complex graph of interrelated entities, in computer memory it is stored as a uniform array that can be efficiently searched and queried. Therefore, triple stores are used when one needs to efficiently perform relatively simple queries. Axioms are statements about ontological classes and instances. Each axiom may be seen as a graph corresponding to one or many RDF triples. Since axioms are more complex than triples, querying them is less efficient. However, with the help of a reasoner, one can execute more

sophisticated DL queries and discover knowledge implicitly contained in the ontology.

Despite the availability of highly developed tools, working with large biomedical ontologies remains challenging. Some of the problems we face are *a)* detection and elimination of errors in OWL documents; *b)* absence of Java virtual machine (JVM) support on some high-performance computing (HPC) platforms; *c)* limited availability of semantic web tools in programming languages such as C++, Python and Perl; *d)* a large footprint of ontology in-memory representation; and *e)* poor parsing and querying performance.

To address these problems we developed *owl_cpp*, a library for parsing, querying, and reasoning with ontologies. Its key features include *a)* strict parsing, detecting errors in OWL documents; *b)* written in standard C++, can be compiled on most platforms; *c)* requires no virtual machine; *d)* possibility of creating efficient APIs for other languages, e.g., Java, Perl, Python; *e)* small memory footprint; and *f)* high performance.

2 Implementation

The following basic operations should be supported by *owl_cpp*: *a)* making a catalog of OWL 2 RDF/XML documents in user-supplied locations, *b)* parsing OWL 2 RDF/XML documents and their imports, *c)* storing and searching resulting RDF triples, *d)* converting the triples into axioms and loading them into a

reasoner, and e) performing description logic queries. This functionality should be implemented along the following guidelines:

- *Correctness and reproducibility* should be verified with extensive unit tests.
- *Strict syntactic verification* should be performed during parsing and axiom generation to prevent possible semantic errors.
- *Error messages* should contain sufficient information to correct the error.
- *Efficiency* should be maximized both in terms of speed and in terms of runtime memory footprint to be able to process large ontologies.
- *Portability* should be maximized by using only standard C++ features.
- *Maintainability* should be maximized by implementing *owl_cpp* as a decoupled modular structure and by utilizing well-established libraries, i.e., the C++ Standard Library and Boost (<http://www.boost.org/>).

Currently, *owl_cpp* is composed of three modules. The **parsing module** is a C++ wrapper for Raptor, a popular C library for RDF parsing [1]. To our knowledge, Raptor is the only C/C++ RDF parser under active development. To parse XML, Raptor uses the SAX interface of Libxml2 library (<http://xmlsoft.org/>). *owl_cpp* reads ontologies only from STL input streams and from filesystem locations specified by the user. Although, by default, Raptor may attempt to fetch ontologies from the internet, this functionality is disabled in the interests of reliability, security, and performance.

The triple store module is responsible for storing, searching, and retrieving of RDF triples. The store internally provides separate containers for namespace URIs, nodes, and triples. The containers keep track of the objects by mapping them against light-weight IDs. Retrieval of an object by such an ID is as efficient as array indexing. Each RDF triple is stored as the three IDs of its corresponding nodes. Although many mentions of the same node may be found in an ontology document, only one instance of each node is kept in the store.

The search of triples is frequently performed and therefore should be highly

optimized. Although a straightforward task, it is made complicated by the number of potential configurations. Depending on the node IDs provided by the user, the triples may be searched in eight different ways: just by the subject, predicate, or object, or by any of their combinations, including the configuration, where no IDs are provided, which should return all triples in the store. Furthermore, as a result of the search, the user may be interested in three types of return values: a complete list of triples that match the provided IDs, only the first triple found, or merely a boolean indicating whether the search was successful. Since for any of the eight search configurations, any of the three types of return values may be required, the total number of possible configurations is 24. Clearly, implementing a separate method for every search configuration would unacceptably clutter the interface. On the other hand, a one-for-all-configurations method would necessarily sacrifice performance. Therefore, in keeping with “you pay only for what you use” principle, the search was implemented as a non-member template function that accepts either `Node_id` or `Blank` types for each of the three terms. The return type is a `boost::iterator_range` (<http://www.boost.org/doc/libs/release/libs/range/index.html>), which can be implicitly converted to a boolean or used for iterating over the matching triples.

The **reasoning module** functionality is performed by FaCT++, which is, to our knowledge, the only open-source C/C++ DL reasoner library [2]. *owl_cpp* passes information from the triple store to FaCT++ by converting triples to axioms. Currently the conversion is done using the Visitor design pattern [3, 4]. DL queries are currently performed directly through the FaCT++ interface with the aid of actor and predicate classes supplied by *owl_cpp*.

3 Results and Discussion

The interface of *owl_cpp* is designed to simplify basic operations with ontologies. For example, the file `ontology.owl` is read into the triple store by calling `load("ontology.owl", store)`. To load the ontology along with its imports, a catalog of ontology IDs and locations should also be provided:

```
load("ontology.owl", store, catalog).
```


Once loaded into the store, the triples can be queried. For example, expression `find_triples(blank, T_rdf_type::id(), T_owl_Class::id(), store)` finds all triples that declare classes. The axioms are copied from the triple store to the FaCT++ reasoning kernel by calling `add(store, kernel)`.

The accuracy of parsing and reasoning of *owl_cpp* was tested with many ontologies. Some of the smaller OWL documents (e.g., several OWL 2 Test Cases, http://owl.semanticweb.org/page/OWL_2_Test_Cases) were incorporated into unit tests, which assert their consistency and satisfiability. Further testing was done during the development of the Ontology of Biological Pathways [5] by executing DL queries formulated by domain experts and comparing the results with ones from Protégé (FaCT++ and HermiT reasoners). The results were always identical.

One of the advantages of *owl_cpp* over other OWL libraries is its ability to discover syntactic errors, thus preventing incorrect semantic interpretation of ontologies. During ontology development in our group, *owl_cpp* has detected inconsistent import statements, undeclared property and annotation predicates, misspelled standard OWL terms, and other problems. Parsing performance of *owl_cpp* was tested using OpenGALEN ontologies version 8 in OWL/RDF format (<http://www.opengalen.org/>). The ontologies occupied 0.5 GB on the hard disk and consisted of 9.7 million triples. Their parsing was successful after correcting several errors, e.g., replacing `owl:propertyChain` terms with `owl:propertyChainAxiom`. The rate of parsing was estimated to be 108 thousand triples per second. The bottleneck of this process appeared to be the insertion of new terms into the triple store. The ways to streamline this step are currently being investigated.

Future development of *owl_cpp* includes the following tasks: *a)* defining high-level C++ APIs for parser, triple store, and reasoner; *b)* designing axiom-based API; *c)* improving readability of error messages; *d)* designing APIs for other programming languages; *e)* introducing support for other OWL 2 syntaxes, e.g., Manchester, Turtle, OWL/XML; and *f)* designing a module for batch execution of OWL 2 Test Cases.

Acknowledgments

The authors would like to thank Dmitri Tsarkov for his help with FaCT++ library. This work was supported by an NIAID-funded R01 (AI077706) and a Burroughs Wellcome Fund Career Award to LGC.

References

1. D Beckett (2001) The Design and Implementation of the Redland RDF Application Framework, in Proc. of the Tenth International World Wide Web Conference.
2. D Tsarkov, I Horrocks (2006) FaCT++ Description Logic Reasoner: System Description, Lecture Notes in Artificial Intelligence, in Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006), Lecture Notes in Artificial Intelligence (Springer), Vol. 4130, pp 292–297.
3. E Gamma, R Helm, R Johnson, J Vlissides (1995) Design Patterns (Addison-Wesley, Boston, MA).
4. A Alexandrescu (2001) Modern C++ design: generic programming and design patterns applied (Addison-Wesley, Boston, MA).
5. AM Masci, MK Levin, A Ruttenberg, LG Cowell (2011) Connecting Ontologies for the Representation of Biological Pathways, in Proc. International Conference on Biomedical Ontology.

Towards Ontologies for ‘Textbook’ of the Future in Obstetrics and Gynecology (OB/Gyn)

Yu Lin¹, Chris Chapman², Erin D. Doelling³, Maya Hammoud^{3,4}

¹Center for Computational Medicine and Bioinformatics, Unit of Laboratory Animal Medicine, and
Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, USA

²Medical School Administration, University of Michigan Medical School, Ann Arbor, MI, USA

³e-Education and Enabling Technologies, Department of Obstetrics and Gynecology Center for Education,
University of Michigan, Ann Arbor, MI, USA

⁴Department of Obstetrics and Gynecology, University of Michigan Medical School, Ann Arbor, MI, USA

Paper not included in Proceedings at the request of the author(s).

Gene Ontology Signatures for Immune Cell-Types Inferred by Gene Expression Analysis

Terrence F. Meehan¹, Christopher J. Mungall²,
Judith A. Blake¹, Alexander D. Diehl³

¹Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA

²Lawrence Berkeley National Laboratory, Berkeley, CA, USA

³University at Buffalo School of Medicine and Biomedical Sciences, Buffalo, NY USA

Abstract. The Cell Ontology (CL) is a candidate OBO Foundry ontology for the representation of *in vivo* cell types. The CL is being revised to include logical definitions for cell types by using terms from other OBO ontologies such as the Gene Ontology (GO). For example, a “T-helper 17 cell (CL:0000899)” is *capable_of* “interleukin-17 production (GO:0032620)”. Computational reasoners exploit these definitions to automate classification and ensure internal consistency in the ontology, and can sometimes expose unexpected but logically sound associations between cell types. These inter-ontology links are typically generated manually using biological knowledge and established literature. We are developing a method to go beyond these sources by integrating ontology and gene expression analysis in order to find “GO term signatures” of cell-types. Using data available from the Immunological Genome project (www.immgen.org), we have generated pair-wise gene expression comparisons between 88 immune cells types represented in the CL. By placing these datasets within an ontological context, GO term signatures are inferred not only for the sampled cell types but also for many ancestral cell types positioned higher in the hierarchy of the CL. Some of these signatures provide biological insights such as defining mature B cells as mature lymphocytes that are *capable_of* MHC class II presentation. This work demonstrates the utility of the CL in gene expression analysis and can be used to semi-automate GO term signatures to cell types.

Keywords: Gene Ontology, Gene Expression Analysis, immune processes

Aligning Research Resource and Researcher Representation: The eagle-i and VIVO Use Case

Stella Mitchell¹, Carlo Torniai², Brian Lowe¹, Jon Corson-Rikert¹, Melanie Wilson²,
Mansoor Ahmed³, Shanshan Chen³, Ying Ding³, Nicholas Rejack⁴, Melissa Haendel²,
the eagle-i Consortium, and the VIVO Collaboration

¹Albert R. Mann Library, Cornell University, Ithaca, NY, USA

²Oregon Health and Science University, Portland, OR, USA

³Indiana University, Bloomington, IN, USA,

⁴University of Florida, Gainesville, FL, USA,

Keywords: expertise, research resources, linked open data, application ontology, researcher profile

1 Introduction

People are a uniquely pervasive nexus, linking the inputs, activities, and outputs of research over time and thus enabling their discovery through contextual relationships with other people, organizations, and events. However, an artificial separation between information about people and information about the research resources they use has evolved in institutional information systems. This lack of interoperability has recently been recognized as impairing the efficiency and effectiveness of research. The impairment stems not only from the overall lack of discoverability, but also from the lack of connectivity between systems for managing information about people and those, if any, that gather and maintain information about research resources. The lack of data integration is compounded when working across institutions, where local needs may dictate inconsistent approaches to data collection, management, and display.

Many institutions are adopting or building researcher profiling systems to encourage collaboration, and these typically encompass publications and grants as the most visible and widely accepted evidence of scientific productivity and expertise. However, few if any, of these systems encompass other products of research or research resources. Research resources can be, however, as compelling evidence of expertise and experience as publications, presentations, and other more commonly tracked outcomes. They provide another vehicle for discovery as well as rich

connection opportunities across the often-hidden research infrastructure of a university or medical school.

The National Center for Research Resources (NCRR) 2009 U24 request for proposals, which funded the eagle-i and VIVO projects, aimed to “develop, enhance, or extend infrastructure for connecting people and resources to facilitate national discovery of individuals and of scientific resources by scientists and students to encourage interdisciplinary collaboration and scientific exchange” [1]. The U24 RFA distinguished projects about research resources from those about the researchers themselves, but eagle-i [2] and VIVO [3] have been strongly encouraged by the NIH to work together to address information technology needs of the research community.

2 VIVO and eagle-i are Ontology-Driven Applications

VIVO and eagle-i store data natively in RDF and use OWL ontologies as primary data models for their respective applications. Both projects have been successful in building flexible and extensible applications leveraging the relationships as much as the type differentiation inherent in ontology class hierarchies. Both offer web-based editing of content in addition to display and search functionalities. As ontology-driven applications that also want to interoperate with the Linked Open Data (LOD) cloud, eagle-i and VIVO bridge some of the gaps dividing the LOD,

RDF-centric world from the more constrained domain of OWL. Both ontologies are designed to support extensive instance data, on the order of tens of millions of statements per university.

The ontologies are also influenced by the high value placed on having eagle-i and VIVO data available as LOD. Our applications support a number of sub-properties to help consumers of linked data to decide what path to follow; we can also facilitate linked data crawling by returning the labels as well as URIs and `rdf:type` of the object individuals of object property statements. The ontologies developed by VIVO and eagle-i aim to facilitate data publishing of. To this end, VIVO uses a local extension process to support individual site needs while still enabling class subsumption to support multi-site search at the VIVO core level. For example, VIVO at Cornell has included research resource types from eagle-i and populated them as a way to have facilities and related services more broadly visible in the Cornell research community. This local extension methodology will likely be applied to eagle-i in the future.

3 Ontology Interoperability

The eagle-i ontology [4] has where possible adopted classes and properties from OBO ontologies that represent entities common in the biomedical context, while the VIVO ontology [5] covers the full range of disciplines represented in major research universities. OBO Foundry principles for ontology development [6] are intended to be orthogonal (non-overlapping) to permit any domain to develop its necessary classes and properties and encourage re-use of classes and properties from other ontologies at points of intersection. The goals are to avoid the ambiguity of closely aligned but distinct class and property definitions, and to avoid the need for mapping by importing and reusing actual URIs from other ontologies. To this end, both projects have aligned their ontologies under the Basic Formal Ontology (BFO) [7] to provide a consistent upper ontology framework guiding the class and property structure. Both use the MIREOT approach [8] to bring together the required ontologies and vocabularies, which reduces complexity and allows interoperability with other systems.

The alignment between VIVO and eagle-i has focused on areas of natural overlap between our respective domains as well as points of intersection with external ontologies of common interest. Some of these classes in common exist in the eagle-i namespace, some in the VIVO namespace, and some are imported and defined in external ontologies such as the Ontology of Clinical Research (OCRe) [9], Ontology for Biomedical Investigations (OBI) [10], Bibliographic Ontology (BIBO) [11] and Friend of a Friend (FOAF) [12]. Since not all of these are within the OBO library of orthogonal ontologies, this has sometimes proven a challenge, as there can be significant overlap and therefore the need to choose and/or map between them. The eagle-i ontology encompasses many more types (classes) than VIVO in order to allow resource contributors to classify data with important nuances. The VIVO ontology has many fewer classes but a relatively larger number of properties to distinguish the many relationships among people and organizations.

4 Challenges and Future Directions

There are areas of overlap between the ontologies where eagle-i uses research terminology while VIVO represents similar concepts in general terms applicable across academia, as illustrated by distinctions between `foaf:Person` and `eagle-i:HomoSapiens`. This raises the question of how using `foaf:Person` to represent a researcher should be reconciled with the notion of `Homo sapiens` as a research subject, e.g. as a source of a biospecimen. In other cases, both ontologies use the same term classified in different ways under the BFO hierarchy based on the anticipated context of use. For example, one ontology represents a service as the offering of a service at a particular point in time, while the other represents ongoing services offered by a core facility. Definitions must also be clarified for properties as well as types: the `hasLocation` property, for example, may relate not only to geographic locations but also to areas of the brain. This is the difference between the use of an ontology in combination with LOD versus simply linking to the same URI based on the class or property label alone -context is

important. Consistent application of ontology terms and properties will be aided by developing a common approach to ontology metadata.

We recognize that controlled vocabularies of terminology (such as MeSH and other UMLS vocabularies) must remain distinct from the eagle-i and VIVO ontologies. However, it is clear that there should be a mechanism for interoperability due to the large amount of research outputs indexed with these vocabularies. Certain subsets, such as the MeSH disease tree, may be imported to promote consistency in annotation of resources. The strategy we chose to follow is to reference external URIs, leveraging the work of the bioinformatics research team under Dr. Moisés Eisenberg at Stony Brook University.

Teams of scientists are now much more common than single scientists in the production of biologically meaningful and clinically consequential breakthroughs [13]. It is common for new grant programs to require multi-institutional and multi-disciplinary participation, as well as to prioritize the participation of underserved populations. Nevertheless, it is still often difficult for investigators to find active researchers and research resources in their field, especially when required to move beyond their own professional circles. We believe that increasing the visibility of the full range of researcher activities and outputs will greatly enhance the discovery potential and promote collaboration. Encoding this knowledge based on the semantic relationships between them lays the groundwork for new methods of inferring expertise based on a wider range of data representing research resources and not just grants and publications as outputs. Of particular interest is mapping clinical experience to medical vocabularies in order to bridge the gap between clinical and basic research domains.

Acknowledgements

VIVO is supported by U24RR029822 and eagle-i by 1U24RR029825 from NIH/NCRR.

References

1. Recovery Act 2009 Limited Competition: Enabling National Networking of Scientists and Resource Discovery (U24), [http:// grants.nih.gov/grants/guide/rfafiles/RFA-RR-09-009.html](http://grants.nih.gov/grants/guide/rfafiles/RFA-RR-09-009.html)
2. eagle-i project, <http://www.eagle-i.org>
3. VIVO project, <http://www.vivoweb.org>
4. Torniai, C., Bashor, T., Bourges-Waldeg, D., Corday, K., Frost, H.R., Johnson, T., Segerdell, E., Shaffer, C.J., Stone, L., Wilson, M.L., Haendel, M.A.: eagle-i: an ontology-driven framework for biomedical resource annotation and discovery. *Bio-Ontologies 2010: Semantic Applications in Life Sciences*, ISMB (2010).
5. Krafft, D.B., Cappadona, N.A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.J., VIVO Collaboration VIVO: Enabling National Networking of Scientists. *Proc WebSci10: Extending the Frontiers of Society OnLine*, April 26-27, 2010, Raleigh, NC: US (2010).
6. Smith, B., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. In: *Nat Biotechnol*, 25 (11), 1251-5. PMID: PMC2814061. (2007)
7. Basic Formal Ontology (BFO), <http://www.ifomis.org/bfo/home>
8. Courtot, M., Gibson, F., and Lister, A. MIREOT: the Minimum Information to Reference an External Ontology Term. *ICBO* (Buffalo, NY) (2009).
9. Ontology of Clinical Research (OCRe), <http://bioportal.bioontology.org/ontologies/39956>
10. Ontology for Biomedical Investigations, <http://obi-ontology.org/>
11. Bibliographic Ontology (BIBO), <http://bibliontology.com/>
12. Friend of a Friend (FOAF), <http://www.foaf-project.org/>
13. Börner K, Contractor N, Falk-Krzesinski HJ, Fiore SM, Hall KL, Keyton J, Spring B, Stokols D, Trochim W, Uzzi B. A multi-level systems perspective for the science of team science. *Science Translational Medicine*, 2(49), 49cm24 (2010).

An Ontology-Based Approach to Linking Model Organisms and Resources to Human Diseases

Christopher J. Mungall¹, David Anderson², Anita Bandrowski³, Brian Canada⁴, Andrew Chatyr-Aryamontri⁵, Keith Cheng⁶, P. Michael Conn⁷, Kara Dolinski⁸, Mark Ellisman³, Janan Eppig⁹, Jeffrey S. Grethe³, Joseph Kemnitz¹⁰, Shawn Iadonato¹¹, Stephen D. Larson³, Charles Magness¹¹, Maryann E. Martone³, Mike Tyers¹², Carlo Torniai⁷, Olga Troyanskaya⁸, Judith Turner¹³, Monte Westerfield¹⁴, Melissa A. Haendel⁷

¹Lawrence Berkeley National Laboratory, Berkeley, CA, USA; ²University of Washington, Seattle, WA, USA;

³University of California, San Diego, CA USA; ⁴University of South Carolina, Beaufort, SC, USA;

⁵University of Edinburgh, UK; ⁶Penn State College of Medicine, Hershey, PA, USA;

⁷Oregon Health & Science University, Portland, OR, USA; ⁸Princeton University, Princeton, NJ, USA;

⁹The Jackson Laboratory, Bar Harbor, ME, USA; ¹⁰University of Wisconsin-Madison, WI, USA;

¹¹Kineta, Inc., Seattle, WA, USA; ¹²Samuel Lunenfeld Research Institute, Toronto, ON, Canada;

¹³TCG, Washinton, DC, USA; ¹⁴University of Oregon, Eugene, OR, USA

Keywords: model organism, phenotype, gene orthology, similarity algorithm

Abstract

The scientific community has invested heavily in the creation of genetically modified organisms, other model systems, and large genetic screens because they greatly inform our understanding of human disease. However, it remains difficult to identify organisms suitable for one's research because information about them is not readily accessible. The initiative to Link Animal Models to Human Disease (LAMHDI; <http://lamhdi.org>) was developed to allow users to search for a diverse set of models of disease using both curated disease-model links and inferred paths based on gene orthology and pathway membership. These inferences are made by traversing connections between records in publicly available data from resources such as the Online Mendelian Inheritance in Man (OMIM), Medical Subject Headings (MeSH), EntrezGene, Homologene, and WikiPathways. This allows researchers to rapidly explore and identify a wide range of model systems, visualizing the multi-step genetic relationship between disease and model. However, if LAMHDI were able to semantically link an organism's phenotypic attributes to diseases, genes, expression profiles, etc, their relevance and utility to a given line of research would be much more greatly illuminated and new novel insights between disease, genetics and phenotype discovered.

The relationship between model systems and disease phenotypes [1] is not straightforward, and bioinformatics tools based on phenotypes have been lacking. Constructed model systems typically only replicate subsets of disease phenotypes, and phenotypes may map to more than one human disease [2]. Classification systems where a model system is listed as a "model of disease X" do not solve these difficulties because the model may only recapitulate one aspect of the disease, but not specifically indicate. Further, different vocabularies are used to describe the phenotypic consequences of mutation in different organisms, and these vocabularies are usually tied to the particular anatomies or physiologies of the organism. The different vocabularies also come from different starting points: clinicians and researchers have different vocabularies. Phenotypes may also occur at different scales, and the relationships among them may not be apparent without additional knowledge. Bioinformatics tools, or even researchers, may not relate the statement 'CA1 dendrites are degenerative' to 'degenerative hippocampus' despite potential scientific correlation. Another challenge is to traverse anatomical structures across species. For example, computers do not know that the human cornea is related to the zebrafish retina because they do not know that in both species these structures are part of the eye, nor do they know that the zebrafish eye is

related to the human eye. When phenotype descriptions are captured using an ontology, algorithms can be written to compare phenotypes computationally across species and scale.

Our previous work has shown that ontological annotation of diseases and phenotypes allows computational comparison of phenotypes across species [3]. To describe phenotypes we composed Description Logic (DL) expressions using a phenotype model and the Web Ontology Language (OWL) [4]. We created bridging ontologies that enhance external ontologies such as the Mammalian Phenotype Ontology (MPO) [5], allowing them to be integrated with other phenotype data [6-8]. We applied these methods to the construction of PKB [9], a neurodegenerative disease phenotype knowledgebase called that utilizes the NIF Standard (NIFSTD) modular collection of ontologies [10] to represent a range of human diseases and animal models spanning multiple anatomical scales, from the molecular and subcellular up to the organismal. We also created an integrative ontology called Uberon [11] to allow cross-species inference. Using these tools and methods, we demonstrated that we could identify organisms with similar phenotypes across anatomical scale, mutations in other alleles of the same gene, other members of a signaling pathway, and orthologous genes and pathway members across species based on the similarity of the phenotypes alone.

Here, we bring together the genetic links currently used within LAMHDI with these ontology-based phenotype similarity methodologies. We also combine orthology and gene-phenotype ontology associations to generate “phenolog” hypotheses, non-obvious linkages between human diseases and phenotypes in model organisms such as mice, worms, yeast, and plants [12]. This approach can suggest new models based on the presence of orthologous genes inside a phenolog cluster. We believe that these ontology-based phenotypic and homology-based techniques will be instrumental in enhancing the LAMHDI portal, suggesting new paths from diseases to models, and assisting in the interpretation of existing paths.

At present, LAMHDI is restricted to providing access to organism strains, and does not include *in vitro* model systems and organismally derived resources such as

biospecimens, cell lines, assays, and reagents. Choosing an appropriate *in vitro* model system to test a given hypothesis is currently not straightforward because these resources are often not linked to diseases, genes, gene expression, or the phenotypes of the organisms from which they are derived. Two projects, the Neuroscience Information Framework (NIF) and eagle-i [13], have built ontologies to catalog and link such resources. The eagle-i system has related specific genotypes of organisms to these *in vitro* models, and has expressly represented anatomical, histological, and pathological attributes of biospecimens and cell lines and tied them into the phenotypic representation of the original organism and its genotype. NIF has focused on collating publicly available resources (materials, software tools, data). The two projects are thus highly complementary.

Navigation of organismal resources (*in vivo* and *in vitro*) also benefited from including gene expression data collected using anatomical ontologies from databases aggregated in NIF. Where genes are preferentially expressed, gene expression profiles can be linked between the current LAMHDI data (genes to disease) with cell lines and tissues. For instance, a researcher looking at a set of genes expressed in a particular brain region may investigate a common mechanism for two diseases if tissue from that brain region is available from a bank that focuses on a different disease. By leveraging the organismal resource component of the eagle-i and NIF systems, and the gene expression component of NIF, LAMHDI will be able to offer many new candidate model systems and make these resources easier to navigate. We discuss the issues of integrating these diverse data and our technical approach to this challenge.

Acknowledgement

LAMHDI is supported by contract HHS N268200800014C from NIH/NCRR.

References

1. Houle, D., Govindaraju, D. R., and Omholt, S. Phenomics: the next challenge.: Nat Rev Genet, 11 (12), 855-66 (2010)
2. Gupta, A., Ludascher, B., and Martone, M.E. BIRN-M: a semantic mediator for solving real-world neuroscience problems. Proceedings of the 2003 ACM SIGMOD international conference on

- Management of data (ACM New York, NY, USA), 678 (2003)
3. Washington*, N.L., Haendel*, M.A. Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7 (11) *Contributed equally (2009)
 4. Mungall, C., Gkoutos, G.V., Washington, N.L., Lewis, S. Representing phenotypes in OWL. In C. Golbreich, A. Kalyanpur, and B. Parsia (eds.): Workshop on OWL: Experiences and Directions (Innsbruck, Austria) (2007)
 5. Smith, C. L, Goldsmith, C. W., and Eppig, J.T.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 6 (1), R7 (2005)
 6. Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E., Ashburner, M.: Integrating phenotype ontologies across multiple species. *Genome Biol*, 11 (1), R2 (2010)
 7. Gkoutos, G.V., Mungall, C.J., Doelken, S., Ashburner, M., Lewis, S., Hancock, J.M., Schofield, P.N., Kohler, S., Robinson, P.N. (2009) Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO. Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)
 8. Hancock, J.M., Mallon, A.M., Beck, T., Gkoutos, G.V., Mungall, C., Schofield, P.N.: Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mamm Genome*, 20 (8), 457-61 (2009)
 9. Maynard, S.M., Mungall, C.J., Lewis, S.E., Martone, M.E.: A knowledge based approach to matching human neurodegenerative disease and associated animal models. *Neuroscience 2010* (230.4; San Diego) (2010)
 10. Bug, W.J., Ascoli, G.A., Grethe, J.S., Gupta, A., Fennema-Notestine, C., Laird, A.R., Larson, S.D., Rubin, D., Shepherd, G.M., Turner, J.A., Martone, M.E.: The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, 6 (3), 175-94 (2008)
 11. Haendel, M.A., Gkoutos, G.V. Lewis, S. Mungall, C.J. Uberon: towards a comprehensive multi-species anatomy ontology.: International Conference on Biomedical Ontologies, Buffalo, NY: Nature Proceedings (2009)
 12. McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., Marcotte, E.M.: Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci U S A*, 107 (14), 6544-9 (2010)
 13. Torniai, C., Bashor, T., Bourges-Waldegg, D., Corday, K., Frost, H.R., Johnson, T., Segerdell, E., Shaffer, C.J., Stone, L., Wilson, M.L., Haendel, M.A.: eagle-i: an ontology-driven framework for biomedical resource annotation and discovery. *Bio-Ontologies 2010: Semantic Applications in Life Sciences*, (Boston, MA) ISMB (2010)

Using RxNorm to Extract Medication Data from Electronic Health Records in the Rochester Epidemiology Project

Jyotishman Pathak¹, Sean P. Murphy¹, Brian N. Willaert¹, Hilal M. Kremers¹,
Christopher G. Chute¹, Barbara P. Yawn², Walter A. Rocca¹

¹Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

²Department of Research, Olmsted Medical Center, Rochester, MN, USA

Abstract. RxNorm is a standardized terminology for clinical drugs developed by the U.S. National Library of Medicine (NLM) in order to facilitate exchange and public availability of medication information. In this study, we evaluate the applicability of RxNorm for representation of medication data from two institutions that are part of the Rochester Epidemiology Project (REP). We detail the researchers' analysis objectives and subsequent requirements for a drug terminology that is comprehensive and easily accessible. We also explore the completeness of mappings between RxNorm and a commercial drug database, Multum, for this sample of REP medications.

1 Introduction

In this study, we develop approaches for structured and unstructured querying of medication data from the out-patient prescription records of two institutions that are part of the Rochester Epidemiology Project (REP [1]). We evaluate RxNorm [2] for standardized data representation from the following aspects: (1) Coverage: What coverage of terms and concepts does RxNorm provide for the REP medication data? (2) Mapping consistency: Are the mappings between RxNorm and external drug databases, such as Multum, accurate and consistent?

2 Background: The Rochester Epidemiology Project

The Rochester Epidemiology Project (REP [3]) is a collaborative effort between several healthcare providers in Olmsted County, MN. It is a medical records-linkage system encompassing the care delivered to all residents of Olmsted County including the Mayo Clinic and Olmsted Medical Center. For this study, we investigate the utility of RxNorm for standardization of REP medication data to facilitate data exchange and interoperability. In particular, based on Mayo's cTAKES

platform [4] and RxNorm, we develop natural language processing (NLP) techniques to standardize medication data extracted from the EHR systems of the two REP providers.

3 Materials

The following materials were used in this study:

- RxNorm January 3, 2011 and November 17, 2008 Full Update release data were used.
- Medication history, as part of out-patient clinical notes (out-patient prescriptions referred to as Orders97), for 212,974 unique individuals was retrieved from the Mayo Clinic's EHR system between January, 2004 and October, 2010. The dataset contained more than 180,000 unique mentions of medications and was retrieved by processing approximately 5 million rows of data from out-patient prescriptions by the cTAKES NLP techniques.
- Out-patient prescription data for 105,151 unique individuals was retrieved from the Olmsted Medical Center EHR system (based on Microsoft® SQL Server), between August, 2002 and November, 2010. The dataset contained 1375 unique mention of medications and corresponding Multum codes extracted via SQL queries.

4 Methods

4.1 Extracting Medication Data via Structured Querying

The Olmsted Medical Center prescription data was extracted from the EHR system (InteGreat IC-Chart) through a scheduled job that queries the data directly from the EHR source database tables. The prescription data, ranging between August, 2002 and November, 2010, was retrieved for 105,151 unique individuals from the primary prescription table, which also included Multum drug codes.

For representing this data using RxNorm, we mapped the Multum codes to RxNorm codes (RxCUIs). Specifically, we loaded the RxNorm January 3, 2011 Full Update release data in a MySQL database, and queried the RXNCONSO table to retrieve the mappings. For example, Hydrogen Peroxide 300 MG/ML Topical Solution with a Multum code=16282 was mapped to RxCUI=91348.

4.2 Extracting Medication Data via Natural Language Processing

To extract drug mentions from Mayo's Orders97 clinical notes data, we first created a dictionary comprising more than 265,000 terms, identified uniquely via a RxCUI code, to assist in the look-up process using RxNorm. This dictionary also comprised of RxNorm codes that were deemed as "obsolete" by the NLM since the patient corpora included more than a decade old drug information. Furthermore, we supplemented the dictionary with 1717 additional terms primarily comprising drug misspellings and abbreviations that were not available in RxNorm. We used open-source Apache Lucene for implementing the dictionary.

For the extraction process, we used Mayo's open-source cTAKES NLP toolkit [4]. In particular, only those terms from the Orders97 data that were composed of seven tokens or less considered for a match in the dictionary. (A

token is considered any text surrounded by blanks with the exception of punctuation characters, which are considered as tokens as well.) Therefore, if a match was discovered for the first term in a drug mention, then all permutations of up to the six remaining tokens were considered for a match. Note that limiting the number of permutations is necessary to minimize the computation time necessary to handle larger drug names. For instance, a ten token drug name would take 362,880 permutation lookup tasks, whereas a seven token name would only take 720 lookups. Each term in the dictionary was run through a tokenizer process that separates the terms into distinct tokens to automatically discover drug mentions and relevant attributes, such as dosage, route, form, frequency, duration, and drug change status.

5 Results

5.1 Using RxNorm for Olmsted Medical Center Data

Since the medication information for Olmsted Medical Center was retrieved using straightforward SQL queries, our focus was primarily on evaluating the mapping between this data, represented using Multum, to RxNorm codes. All of the 1375 unique mention of medications and corresponding Multum codes in the dataset were mapped to RxNorm by querying the RXNCONSO table from the RxNorm release. In order to check accuracy, we randomly selected and manually reviewed (by an experienced pharmacist) the 500 drug mentions of top 50 frequently administered drugs, specifically focusing on medication names and their descriptions. We found that the entire set of mappings for the 500 drug mentions were accurate, thereby validating that the curation of mappings, at least between RxNorm and Multum, done by the RxNorm curators is of high quality and consistency.

RxNorm Release	Unique terms identified by cTAKES NLP medication pipeline	Unique RxCUIs identified	Unique terms with RxCUIs	Unique terms without RxCUIs
Nov. 2008	181,722	7,908	114,653	67,069
Jan. 2011	181,727	8,058	135,988	45,739

Table 1. RxNorm results for Orders97 dataset processing

5.2 Using RxNorm for Mayo Clinic Data

As mentioned above, for Mayo's Orders97 data, we extracted medication information for 212,974 unique individuals using the cTAKES NLP process, and represented it using both the November 2008 and January 2011 releases of RxNorm (both versions were used to facilitate a comparative analysis). In particular, we analyzed 4,964,022 rows representing medication mentions for 212,974 different individuals in the Orders97 data (there were one or more mentions on each row). 181,722 and 181,727 unique terms for 2008 and 2011 RxNorm releases, respectively, were identified as valid medication "related" mentions by the cTAKES pipeline. From this, 7,908 and 8,058 unique RxNorm concept codes were identified for the 2008 and 2011 RxNorm releases, respectively.

DTaP/IPV/Hib Vaccine	Inactivated polio vaccine
Haemophilus Influenzae Type b (Hib) Vaccine	MMRV (measles, mumps, rubella varicella) vaccine
H1N1 swine flu vaccine	Typhoid vaccine
Meningococcal Vaccine	Hepatitis A vaccine

Table 2. Sample list of vaccines

6 Discussion

RxNorm with its vastness provides a comprehensive coverage of drugs and medications. However, in this study, we found that coverage for vaccines require further improvement. Table 2 shows a sample list of vaccines for which no corresponding RxNorm codes were discovered while processing the Orders97 data (this list was manually identified from Orders97). Specifically, we identified 103 distinct vaccine related terms in our Orders97 dataset, of which 35 had recognizable terms in RxNorm, although variations of those terms were missing from RxNorm. For example, the text span "H1N1" produced no hits, but the text span "Influenza A (H1N1) Vaccine 2009" had a corresponding RxCUI. Furthermore, terms such as "influenza H1N1 vaccine" which were not present in the 2008 release of RxNorm, but present in the updated 2011 release. This leads us to hypothesize that coverage for vaccines will potentially improve in future RxNorm releases. In addition to above, our investigation also highlighted that drug and medication terms commonly occurring in the Orders97

dataset which were represented using abbreviations, hyphenations, or aliases (e.g., "vit. b3" to represent "Vitamin B3") did not have a corresponding RxNorm code.

RxNorm also includes terms such as air and water, and forms, such as cream, powder and oil, and ingredient mentions that are associated with minerals or common elements, such as calcium, glucose and magnesium. Such terms tend to create noise when extracting drug content from unstructured text. Therefore, we compiled a list of such terms to prevent the cTAKES NLP pipeline from propagating the annotated named entities in this group, which lead to significant performance improvement in named entity detection.

Limitations and Future Work. In this study, we limited our investigation to analysis of outpatient medication data from only 2 institutions. In future, we plan to incorporate in-patient medication data and other institutions as well. Furthermore, an important requirement for REP is classification and categorization of RxNorm coded medication data using standardized drug classifications, such as NDF-RT. To this end, we are currently incorporating recently released NLM's NDF-RT API [5] within the cTAKES NLP pipeline for mapping and classifying RxNorm coded data from REP to NDF-RT drug classes.

Acknowledgments

Research funded in part by the Rochester Epidemiology Project (AG034676-45) and Mayo Clinic Early Career Award to the first author.

References

1. St. Sauver, J., et al., *Use of a Medical Records Linkage System to Enumerate a Dynamic Population Over Time*. Am. J. Epidemiol 2011.
2. Liu, S., et al., *RxNorm: Prescription for Electronic Drug Information Exchange*. IT Professional, 2005. 7(5): p. 17-23.
3. Melton, L., *History of the Rochester Epidemiology Project*. Mayo Clinic Proceedings, 1996. 71(3): p. 266-274.
4. Savova, G., et al., *Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications*. JAMIA, 2010. 17(5): p. 507-513.
5. *National Library of Medicine NDF-RT Web Services API. Last updated on: February 16th, 2011*. <http://rxnav.nlm.nih.gov/NdfrtAPI.html>.

Towards Desiderata for Provenance Ontologies in Biomedicine

Satya S. Sahoo

Division of Medical Informatics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

Abstract. Provenance is essential metadata in biomedicine to verify data quality, ensure reproducibility of published data, validate experiment protocols, and compute trust of scientific results. Using the requirements identified by the W3C Provenance Incubator Group, seven desired attributes are defined to create an evaluation framework for ten provenance ontologies in biomedicine. Three ontologies, the Experimental Factor Ontology, the Parasite Experiment Ontology, and Ontology for Clinical Research are found to be fully compliant with the desiderata.

1 Introduction

Provenance, derived from the French word *provenir* meaning “to come from” is critical contextual metadata in biomedicine to validate data quality, verify integrity of experiment processes, and compute trust [1] [2]. Provenance metadata is also required for ensuring reproducibility of published scientific results, objective comparison of datasets produced by different research groups, effective biomedical data integration (in form of contextual metadata) [2]. The objective of this study is to propose a set of desired attributes for a provenance ontology in biomedicine that addresses some of the requirements identified by the W3C Provenance Incubator Group and also some of the OBO Foundry principles, and review a list of existing provenance ontologies using the desired attributes as a frame of reference.

Existing work in identifying desired qualities of biomedical ontologies include, evaluation framework for controlled medical vocabularies [3], disease ontologies [4], and the OBO foundry principles (http://www.obo.foundry.org/wiki/index.php/OBO_Foundry_Principles).

We derive some of the attributes used in this study from the existing work, but in addition use the “requirements for provenance” identified by the W3C Provenance Incubator Group [5], which are desired attributes relevant for provenance management in biomedicine.

2 Methods

First, a set of existing biomedical ontologies suitable for modeling provenance information are selected. Next, we define the comparison framework by identifying the characteristics that will facilitate provenance management in biomedicine. The selected ontologies are analyzed with respect to the set of desired attributes and the results are represented in an attribute versus provenance ontology matrix.

3 Candidate Ontologies

The National Center for Biomedical Ontologies (NCBO) was the primary source of candidate ontologies, where the “Experimental Conditions” category was used to identify relevant ontologies. We also used our knowledge of other ontologies modeling biomedical provenance to identify additional ontologies. The ten selected ontologies are briefly discussed below:

1. **ProPreO Ontology.** Ontology for modeling the proteomics analysis pipeline as part of the biomedical glycoproteomics project at the University of Georgia.
2. **Ontology for Biomedical Investigations (OBI).** One of the largest and most comprehensive provenance ontologies covering more than 18 communities, including proteomics, transcriptomics, imaging, and toxigenomics.

3. **Experiment Factor Ontology (EFO).** Ontology developed by the European Bioinformatics Institute (EBI) to model the experimental factors associated with the ArrayExpress database of gene expression and related microarray datasets.
4. **Experiment Conditions Ontology (XCO).** This ontology is one of the three ontologies created for phenotype measurement data.
5. **Biological Imaging Methods (FBbi).** The ontology models information about the sample preparation methods, imaging process, and visualization techniques used in biomedical imaging that influence the quality and subsequent interpretation of the images.
6. **Parasite Experiment Ontology (PEO).** PEO extends the Provenir top domain ontology for provenance (which in turn uses some Basic Formal Ontology (BFO) classes) to model provenance information of bench biological processes used in human pathogen research.
7. **Ontology for Clinical Research (OCRe).** The OCRe ontology models the provenance associated with human studies, both interventional and observational, which span the design phase, study execution phase, and analysis phases.
8. **Cardiac Electrophysiology Ontology (EP):** The Cardio Vascular Research Grid (CVRG) has developed the Cardiac Electrophysiology Ontology to represent metadata describing the experimental conditions for cardio vascular research.
9. **Neural ElectroMagnetic Ontology (NEMO).** The ontology aims to represent the provenance information associated with Electro-encephalography (EEG) and Magnetoencephalography (MEG) data to facilitate collection, sharing, and mining of brain electromagnetic data.
10. **SWAN Provenance Authoring and Versioning (PAV) Ontology.** The ontology was developed as part of the Semantic Web Applications in Neuromedicine (SWAN) project and represents the derivation, authoring, and

versioning information of biological resources.

4 Desirable Features

We identify seven desirable features for provenance ontologies in biomedicine based on both the requirements identified by the W3C Provenance Incubator Group [5] and the ten OBO foundry principles. The W3C Provenance XG identified a number of requirements for provenance along three dimensions, namely (a) content, (b) usage, and (c) management [5]. In the context of provenance ontologies in biomedicine, we believe provenance *interoperability*, *accessibility*, *entailment*, *versioning*, and *understanding* for end users (to enable use of provenance in applications) are the essential desired attributes [5].

1. **Open source without intellectual property restrictions.** Both the W3C Provenance XG accessibility dimension and the OBO foundry principle#1 recommend that the provenance ontology should be freely available without usage restriction or subject to payment of fee.
2. **Facilitating provenance interoperability** by extending upper level ontologies for creation of domain-specific provenance ontologies.
3. **Well-defined representation format.** Corresponding to the Provenance XG requirement for supporting entailment and OBO foundry principle #2, the provenance ontologies need to be available in a standard representation format to support entailments by available reasoning tools.
4. **Usability in real world applications.** This requirement reflects the “understanding” category of the provenance dimensions defined by the Provenance XG, which facilitates the use of provenance in end-user applications.
5. **Continued development and maintenance.** An important challenge for the biomedical ontology community is ensuring the continued development and maintenance of ontologies, as reflected in the OBO foundry principle#4. Hence, provenance ontologies should continue to

be developed and modified as the requirements of provenance users evolves.

6. **Re-use of existing ontologies.** Corresponding to the OBO foundry principle#5, provenance ontologies in biomedicine need to re-use terms from the large number of biomedical ontologies already created by the community.
7. **Explicit support for versioning.** Versioning information is an important aspect of provenance management as identified by the Provenance XG, hence provenance ontologies themselves need to include explicit support for versioning information.

A framework for evaluating provenance ontologies in biomedicine

Similar to the framework for comparing disease ontologies [4], the seven desired characteristics are not given equal weights. We identify some of the attributes reflecting the requirements of the Provenance XG and some OBO foundry principles to have higher importance as compared to others. We give a maximum weight of 5 to attributes numbered (2), (3), (4), and (7); followed by the weight of 3 to attribute (1); and finally weight of 1 to attributes (5) and (6). The ten provenance ontologies reviewed in the paper are assigned a score of 1 (for full support to a given desired attribute), 0 (for no support for the desired attribute), and a discrete value between 0 and 1 depending on the level of support for the desired attribute.

5 Results

Table 1 represents the results of our evaluation. The findings demonstrate that many existing biomedical provenance ontologies, EFO, PEO, and OCRe, fully support the desired properties identified in the evaluation framework. This is an encouraging trend for the biomedical provenance community and needs to be incorporated in other ontologies, such as the SWAN PAV and XCO, which scored less than 50%.

6 Discussion

The primary areas of concern for provenance ontologies are the support for interoperability, which is either partially (for ProPreO) or not supported at all (EP, PAV, XCO, and FBbi). The use of upper-level ontologies, such as BFO or the Provenir top domain provenance ontology [1] are needed to support consistent modeling, use of ontology design patterns and best practices. The re-use of existing ontologies in the creation of new ontologies has been a focus of continued concern for the OBO Foundry. But, five provenance ontologies are found to have no support for re-use of existing ontologies (ProPreO, XCO, FBbi, NEMO, and SWAN PAV). Hence, it is essential for the provenance ontologies community to ensure maximum re-use of existing ontology terms in development of new ontologies.

	Wt.	ProPreO	OBI	EFO	XCO	FBbi	PEO	OCRe	EP	NEMO	PAV
Open source	3	1	1	1	1	1	1	1	1	1	1
Inter-operability	5	0.5	1	1	0	0	1	1	0	1	0
Standard format	5	1	1	1	1	1	1	1	1	1	1
Understanding	5	1	0.5	1	0	0	1	1	0	1	0.5
Continued development	1	0	1	1	1	1	1	1	0	1	0
Re-use ontologies	1	0	1	1	0	0	1	1	1	0	0
Versioning	5	1	1	1	0	1	1	1	1	1	0
Total Score	25	82%	90%	100%	36%	56%	100%	100%	56%	96%	34%

Table 1. The desiderata applied to provenance ontologies in biomedicine

ProPreO	Proteomics data and process provenance, http://bioportal.bioontology.org/ontologies/13386
OBI	Ontology for Biomedical Investigations, http://bioportal.bioontology.org/ontologies/44899
EFO	Experimental Factor Ontology, http://bioportal.bioontology.org/ontologies/39885
XCO	Experimental Conditions Ontology, http://bioportal.bioontology.org/ontologies/45362
FBbi	Biological imaging methods, http://bioportal.bioontology.org/ontologies/45253
PEO	Parasite Experiment Ontology, http://bioportal.bioontology.org/ontologies/42093
OCRe	Ontology for Clinical Research, http://bioportal.bioontology.org/ontologies/44778
EP	Cardiac Electrophysiology Ontology, http://bioportal.bioontology.org/ontologies/39038
NEMO	Neural ElectroMagnetic Ontologies, http://bioportal.bioontology.org/ontologies/45141
PAV	Provenance, Authoring and Versioning Ontology, http://swan.mindinformatics.org/ontologies/1.2/pav.owl

Table 2. List of biomedical provenance ontologies reviewed in this work

7 Conclusions

We use the W3C Provenance Incubator Group recommendations and OBO foundry principles to define a framework of desired attributes for biomedical provenance ontologies. We use the framework to evaluate ten ontologies and find that a majority of ontologies are compliant.

References

1. S. S. Sahoo, "Semantic Provenance: Modeling, Querying, and Application in Scientific Discovery," Ph.D., Computer Science and Engineering Department, Wright State University, 2010.
2. C. Goble, "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics," in *Workshop on Data Derivation and Provenance*, Chicago, 2002.
3. J. J. Cimino, "Desiderata for controlled medical vocabularies in the twenty-first century.," *Methods Inf Med*, vol. 37, pp. 394-403, 1998.
4. Bodenreider O and Burgun, "Desiderata for an ontology of diseases for the annotation of biological datasets.," in *First International Conference on Biomedical Ontology (ICBO 2009)*, NY, USA, 2009, pp. 39-42.
5. *W3C Provenance Incubator Group Wiki*. Available: http://www.w3.org/2005/Incubator/prov/wiki/Main_Page

Modeling Issues and Solutions: Building a Taxonomy from a Biology Textbook

A. Patrice Seyed¹, John Pacheco², Andrew Goldenfranz³, Vinay Chaudhri²

¹Department of Computer Science and Engineering, University at Buffalo, NY, USA

²SRI International, Menlo Park, CA, USA; ³Fremont Union High School District, CA, USA

1 Introduction

Our task is to create a taxonomy from an AP Biology textbook's glossary terms [1] for Project Halo [2]. Project Halo's goal is to build a reasoning system capable of answering novel questions and solving advanced problems in a broad range of scientific disciplines. In support of this goal, the resulting taxonomy is to be used as a foundation for translating passages within the biology textbook into logical formulas on which a reasoning system will operate.

In order to bootstrap our task, we imported ~2400 glossary terms and definition strings from the textbook's electronic glossary into Collaborative Protégé in OWL format, as classes and comment strings. Our team consisted of biologists and KR specialists. We took an iterative approach, where the biologists of our team attempted initial classifications, restricted to the *subclass-of* relation, and were encouraged to add additional classes when they deemed it appropriate. As they proceeded, modeling issues were identified and discussed during workgroup sessions. The issues and the solutions that were implemented are as follows.

2 Results

2.1 Entity/Role Dichotomy

Initially the biologists of our team encoded classes for organic molecules both on the basis of their structure and on the basis of their function (see Figure 1). For instance, proteins and steroids are defined by their chemical composition. In



Figure 1. Naïve Classification of Steroids and Hormones

contrast, hormones are defined by the function they perform, and there is overlap between **Steroid** and **Hormone**, i.e. some hormones are steroids while others are proteins.

As a solution, we define hormones as roles that certain chemicals play. However, **Steroid-Hormone** remains a class in the taxonomy, which represents a useful class of biologists' intuitive thinking (see Figures 2 and 3).

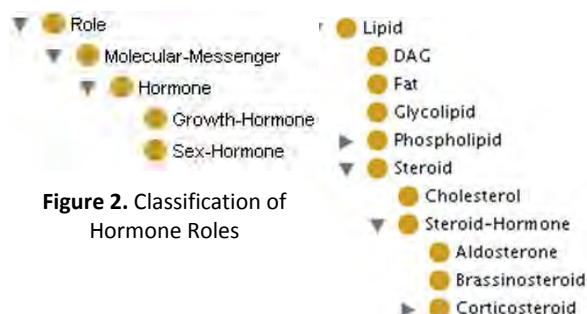


Figure 2. Classification of Hormone Roles

Figure 3. Classification of Steroids Based on Molecular Structure

2.2 Linnaean Classification

The biologists of our team wanted to classify the different kingdoms under the class **Kingdom** (see Figure 4). However, there are 5 instances of **Kingdom** (in the context of a U.S.-based textbook).



Figure 4. Naïve Classification of Kingdoms

As a solution, we represent Linnaean taxonomy under organism, and used common English names for simplicity (see Figure 5). For example, "Cow is an Animal" is clearer than "Cow is an Animalia":



Figure 5. Classification of Organisms



Figure 6. Treatment of Classification Units

As potential refinements, we can add the Latin-named classes as instances of their classification unit (see Figure 6). For example, **Animalia** is an instance of **Kingdom**, and **Chordata** is an instance of **Phylum**. As yet another approach (not pictured), we can treat classification units as meta-classes. For example, **Chordate** is an instance of the meta-class **Phylum**, and **Animal** is an instance of the meta-class **Kingdom**.

2.3 Entity/Process Dichotomy

The biologists of our team wanted to classify **Light-Microscope** under the subclass **Technology** (see Figure 7). They also wanted to classify **Technology** under the subclass **Inquiry**. These two uses of the term 'Technology' refer to two different senses. The glossary definition for Technology is "The application of scientific knowledge for a specific purpose, often involving industry or commerce but also including uses in basic research."

Our solution in this case was to refactor the taxonomy (see Figures 8 and 9). We noted that some terms of the glossary are polysemous. Definitions including "also" were a clear indicator of this. For example, Wild Type is "An individual with the phenotype most commonly observed in natural populations; also refers to the phenotype itself."



Figure 8. Classification of Processes



Figure 7. Naïve Classification of Technology



Figure 9. Classification of Artifacts

2.4 Classifying Areas of Research

The initial tendency was to classify areas of research (e.g. **Genetics**, **Anatomy**, **Ecology**) under **Inquiry**. Areas of research are complex social entities, involving research activities and educational institutions constituted of departments, faculty members, programs and curricula. However, the definitions of the terms for each area of research is prefixed by "the scientific study of". Given the commitment to processes, their classification under **Inquiry** is appropriate.

2.5 Subclass/Subprocess Dichotomy

There was a strong initial tendency to use the hierarchy to organize sub-parts or sub-processes (see Figure 10). For example **Telophase** is a subclass of **Mitosis**, instead of a sub-process.

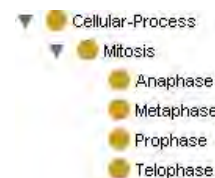


Figure 10. Naïve Classification of Processes

Our solution was to move parts or sub-processes to appropriate locations whenever they are found (see Figure 11). During workgroup sessions, we reinforced how to use the subclass of relationship consistently.

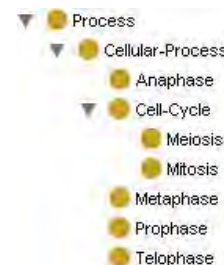


Figure 11. Refactoring of Processes

3 Conclusions

Initially, the biologists of our team relied on prior knowledge and definitions for organizing the class hierarchy, and the classes were treated as organizational "buckets". Ontological principles were iteratively applied to identify modeling issues and provide a foundation for the taxonomy building process.

After following this process for several workgroup sessions, the biologists had a much better sense for these types of modeling issues, and hence were more effective in continuing the taxonomy building process. These lessons and resulting taxonomy can help AURA better answer “What is” questions. Furthermore, these lessons can be applied to other ontologies, although it may depend on what formalism is used (e.g., for dealing with meta-classes).

Acknowledgements

This work was funded by Vulcan Inc.

References

1. Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, Robert B. Jackson. (2010) Campbell Biology, Pearson Publishing.
2. Gunning D. et. al. (2010) Project Halo Update – Progress Toward Digital Aristotle, AI Magazine, Fall 2010, 33-58.

Biomedical Analyses: OWL Model Based Edition

Pierre-Yves Vandenbussche^{1,2}, Ferdinand Dhombres¹, Sylvie Cormont³, Jean Charlet^{1,3}, Eric Lepage³

¹INSERM UMRS 872 ÉQ.20, Paris, France

²Mondeca, Paris, France

³AP-HP Assistance Publique – Hôpitaux de Paris, Paris, France

Abstract. Background and Objectives. The Assistance Publique Hôpitaux de Paris (Public hospital of Paris and its suburbs; APHP) developed a biology dictionary independent from laboratory management systems (LMS). This dictionary is interfaced with the international nomenclature Logical Observation Identifiers Names and Codes (LOINC), and developed in collaboration with experts from all biological disciplines. We aim to establish a platform for publishing and maintaining the APHP laboratory data dictionary, which can satisfy both the requirements concerning the controlled vocabulary and those related to maintenance processes and distribution. **Material and Methods.** Data complexity and data volume show the need to establish a platform dedicated to the terminology management. This replaces the use of a spreadsheet tool that might show weaknesses. After describing the dictionary, we identify requirements for the nomenclature management, and the inadequacy of existing software. Our method is based on the design of a OWL hub meta-model supervising organization systems. **Results.** We describe how the modeling, data migration and integration/verification steps in the new tool were used to meet these requirements. The core of our work is based on the modeling effort which integrates multiple dimensions: (i) interoperability regarding data exchange standards, and (ii) dictionary evolution. This model has been implemented in the APHP context. Structuring data representation has led to a significant data quality improvement.

1 Introduction

One of the projects of the Assistance Publique-Hôpitaux de Paris ¹ (AP-HP) new information system is to acquire a biological analysis dictionary (AnaBio) common to the whole production chain: prescription, analysis processing in the laboratory management systems (LMS) and transmission of the result. Useable by all the LMS, active in the 45 hospitals of the institution spread into 165 laboratories, the repository on which is based this dictionary should ideally remain independent of these tool constraints. The dictionary achieved offers the managerial flexibility necessary to its daily use while maintaining the semantic interoperability with other international health organisms through its alignment with LOINC (Logical Observation

Identifier Names and Codes) [5]. This is also the choice made by other hospitals showing a more or less complete interfacing with LOINC [4]. With this regard, the biomedical analysis dictionary is a perfect example of local terminology interfaced with a reference terminology [3]. This dictionary is also linked to adjunct data such as the list of hospital user facilities as well as their contacts.

In this project, we aim to implement an editing and maintenance platform of the AP-HP biomedical analysis dictionary that may please the requirements related to both the repository and maintenance and diffusion processes. The current repository management software (spreadsheet) shows some adaptation limitations to the dictionary requirements and perspectives. To achieve this goal, our efforts focus on an ontology model definition which supports the representation of the terminologies used and also of adjunct knowledge. This modeling effort must permit the multi-terminological representation, terminology update (e.g. half-yearly LOINC

¹ Assistance Publique – Hôpitaux de Paris. Public hospital of Paris and its suburbs. AP-HP is the largest hospital system in Europe and one of the largest in the world.

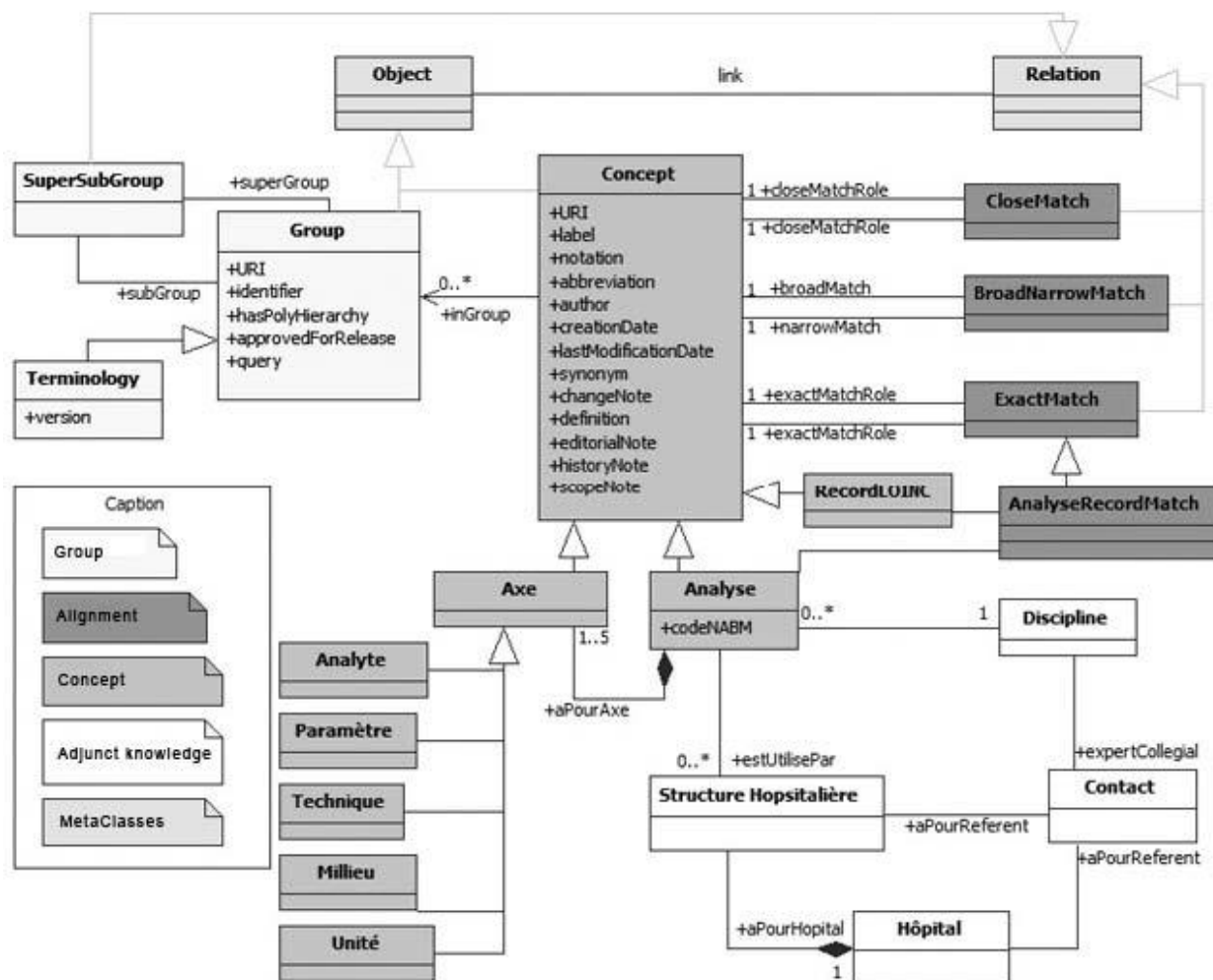


Figure 1. Simplified model of UML class of the AnaBio Ontology model.

update), and possibility to store the translation of LOINC. The data formalization must also allow improving data quality.

2 Method and Results

Our method apprehends the diversity of the terminologies pattern expressivity by defining a unique ontology representation model. This unique model presents the advantage to integrate the different terminologies within a single server and thus, to allow the editing of these repositories. The new platform implementation requires the transmission of semi-structured data to a structured model representation which relies on the knowledge engineering techniques [1]. This change implies modeling, data migration, and integration/validation steps.

The modeling task is conducted in close collaboration with the Terminology Maintenance Unit. This collaboration aims understanding the usefulness of each component and apprehending the new needs which impact the model to design. For example, the addition of status properties allows a better traceability of the components over the time. The structuring of the spreadsheet existing in tabs and columns constitutes a first organization step, discipline understanding, and data exchange need integration. The model is enough generic to represent any type of terminology including LOINC, AnaBio and also future resources that will be useful to improve the interoperability, such as SNOMED-CT. However, this model remains extensible to consider the particularities of each terminology (for example the NABM codes for biomedical

analysis results) but also to link the AnaBio dictionary to the adjunct knowledge such as hospital facilities, contacts, etc. Our approach does not pretend to define an ex nihilo model but wants to be a good practice paradigm for the controlled vocabulary representation. Our method uses and extends parts of modeling in existing norms and standards such as SKOS (Simple Knowledge Organization System) [6] and BS 8723 (British Standard 8723) [2]. The OWL language and its expressivity in description logic are used to describe our model [7] which is presented in Figure 1. Parallel to the model design, the task of **data migration** begins. This requires transforming the entire spreadsheet data to allow their integration and the conformity alignment with the new formal model. Modeling and data migration stages allow an iterative refinement work. To be validated, modeling suggestions are imported into the platform with the migrated data. During the **validation** task, the terminology maintenance team validate, correct and improve the ontology. After 6 validation cycles, the platform deployment in the production environment intervenes.

The improvement in data quality included in the AnaBio dictionary is a major point of the results obtained with this project. The transition from semi-structured data (spreadsheet) to structured data (by the formal model) has forced the correction of data considered as incoherent. Most of these inconsistencies were differences of breakage, spelling or absence of normalization of a value which should be identical. For example, “cysterceques”, “Cysterceque” and “Cysticerques anticorps” will be change to “Cysticerques anticorps”. These corrections aim to improve data quality.

3 Conclusion

This ontology model is a particularly suitable and scalable solution for the needs of a terminology used daily. Contrarily to a XLS file which is constraint by its structure, this model can be extended without impacting the controls, exports and statistics already

implemented. Its generic nature allows the future integration of other terminologies. It also allows the definition of restriction, inference and control rules through its formal definition. The platform implementation to manage biomedical analyses and their associated data highlights some issues which were hidden until then. More than 10% of original data have been corrected during this project. The implemented solution automates and integrates a large number of tasks (automatic index creation, input constraint control defined in the model, etc.), releasing the team in charge of the AnaBio dictionary from proceedings outside of their expertise.

References

1. AussenacGilles, N.: Méthodes ascendantes pour l'ingénierie des connaissances. Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France (décembre 2005), <ftp://ftp.irit.fr/IRIT/CSC/HDR-Aussenac13fev-06.pdf>
2. BS8723: Structured vocabularies for information retrieval, part 4: Interoperability between vocabularies, (2008)
3. Daniel, C., Buemi, A., Mazuel, L., Ouagne, D., Charlet, J.: Functional requirements of terminology services for coupling interface terminologies to reference terminologies. In: Studies in health technology and informatics. vol. 150, p. 205 (2009)
4. Lin, M., Vreeman, D., McDonald, C., Huf, S.: A characterization of local loinc mapping for laboratory tests in three large institutions. *Methods Inf Med* 2010, 49 (2009)
5. McDonald, C., Huf, S., Suico, J., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., et al.: Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry* 49(4), 624 (2003)
6. Miles, A.: Skos: requirements for standardization. In: DC-2006: Proceedings of the International Conference on Dublin Core and Metadata Applications. pp. 55 64 (2006)
7. Vandenbussche, P.Y., Charlet, J.: Méta-modèle général de description de ressources terminologiques et ontologiques. In: Ingénierie de la Connaissance (IC) (2009)

Ontobee: A Linked Data Server and Browser for Ontology Terms

Zuoshuang Xiang¹, Chris Mungall², Alan Ruttenberg³, Yongqun He¹

¹University of Michigan Medical School, Ann Arbor, MI, USA

²Lawrence Berkeley National Laboratory, Berkeley, CA, USA

³Science Commons, Cambridge, MA, USA

Abstract. The Linking Open Data (LOD) community has been working to extend the Web with a data commons by publishing various open datasets as RDF on the Web. To support this effort, we developed Ontobee (<http://www.ontobee.org/>), a web system aimed to serve as a linked data server and browser specifically targeted for ontology terms. Ontobee combines two basic features for one specific ontology term: (1) a user-friendly web HTML interface for displaying the details and its hierarchy of a specific ontology term; and (2) a RDF/XML source code for this ontology term corresponding to the HTML web page, which can be accessed by Semantic Web applications. Ontobee provides an efficient and publicly available method to promote ontology sharing, interoperability, data integration, and Semantic Web applications.

Keywords: Ontobee, ontology visualization, Semantic Web

1 Introduction

Biomedical ontologies are consensus-based controlled biomedical vocabularies of terms and relations with associated definitions, which are logically formulated to promote automated reasoning. The Semantic Web is a web of data that allows machines to understand the meaning – or “semantics” – of information on the World Wide Web. To ensure that computers can understand the semantics of terms, machine-readable ontologies play a central role in Semantic Web development. However, how to present the meaning of ontology terms by its URI is still a challenge. Most ontology URIs do not point to real web pages. While some URIs point to specific pages, the pages shown are often in: (a) pure HTML format (*e.g.*, the Ontology Lookup Service or OLS [1] and NCBO BioPortal) without a RDF source code, or (b) RDF format (*e.g.*, many ontology term URLs such as:

http://purl.org/obo/owl/SO#SO_0001411

which contains the whole ontology instead of individual terms. In both cases these pages do not efficiently support the Semantic Web.

The objective of the Linking Open Data (LOD) community is to extend the Web with a data commons by publishing various open

datasets as RDF on the Web. These RDF links between data items can come from different data sources and accessed anywhere through the web. All of the sources on these LOD diagrams are open data. To support LOD, one basic requirement is to map individual ontology terms to real RDF files through the Web. Many LOD browsers are available *e.g.* Ontology-browser, VisiNav, and others listed at <http://x.co/XjWT>. However, these programs focus on RDF semantic data structure browsing without returning fragmented ontology term information in RDF format.

In this report, we present our development of Ontobee (<http://www.ontobee.org/>), a linked data server and browser for linking and browsing individual ontology terms. Ontobee integrates both ontology content visualization through HTML web pages and Semantic Web content transferring through RDF/XML format.

2 Methods

2.1 Server and Programming Tools

The Ontobee application server is implemented using one Dell Poweredge 2580 server. This server runs the Red Hat Linux operating system (Red Hat Enterprise Linux 5 server).

The open source software program, Apache HTTP Server, is installed as the HTTP application server. PHP is used as programming languages in the web application server.

2.2 Ontology Source Used in Ontobee

The primary ontology source comes from Neurocommons (<http://neurocommons.org>), a knowledge base containing all the ontologies in OBO foundry [2]. Ontobee also supports ontologies from other data sources. For example, Ontobee also maintains a Virtuoso RDF store that contains several ontologies such as the Vaccine Ontology (VO) and the Adverse Event Ontology (AEO).

2.3 Visualization of Ontology Hierarchies and Individual Terms

A software program based on PHP and SPARQL was developed to visualize the hierarchical tree and individual terms from a specific ontology. The ontology tree is assembled from the results of a set of SPARQL queries against the RDF stores. The 'transitive' option in the Virtuoso SPARQL engine was used to minimize the number of SPARQL queries. For individual terms, a HTML page is assembled from the results of another set of SPARQL queries against the RDF stores. The 'transitive' and the 'CBD' (*i.e.*, Concise Bounded Description) options in the Virtuoso SPARQL engine were used to minimize the number of SPARQL queries. In both cases, PHP interacts with the RDF stores by sending SPARQL queries to the stores. The stores then return the results back to PHP scripts in JSON (JavaScript Object Notation) format. The PHP scripts decode information from the JSON formatted results and format it into a user-friendly HTML page.

2.4 RDF Output of Ontology Term Information

In addition to a HTML file, another file is generated to output the ontology term content in OWL (RDF/XML) format. This OWL format is the default page format for the term URI. A link to the HTML file is provided in the OWL format through XSLT (Extensible Stylesheet Language Transformations; <http://www.w3.org/TR/xslt>). When the term URI was visited by

using a web browser, both the OWL file and the HTML content will be retrieved but only the HTML content will be shown in the browser for easy reading. A user can easily access the OWL content by checking the source code of the HTML page. Alternatively, without a human visit to the HTML page, the OWL file can be retrieved by a web application or a Semantic Web system.

3 Results

3.1 Software Design and Statistics

The Ontobee software integrates two basic features: ontology content visualization, and RDF content sharing (Fig. 1). Basically, a user can use a web browser to query Ontobee web page. Ontobee will then issue SPARQL queries to a RDF triple store. The results of the SPARQL queries are returned in JSON format and processed by Ontobee PHP scripts to form a user-friendly HTML page. Simultaneously, RAW OWL output was also returned and used to generate machine-readable RDF output file. A web application can be developed to access the RDF output file directly without access using a web browser.

Currently Ontobee provides access to 88 ontologies, the majority of which come from the OBO Library. These ontologies cover different domains, such as anatomy, health, and experiments. In total, 759,003 ontology terms from 88 ontologies are currently covered.

3.2 Ontobee Features

Fig. 2 is an example of using Ontobee for the term 'vaccine' (VO_0000001) in the Vaccine Ontology (VO). The HTML web page displays the term information (*e.g.*, term equivalents and class hierarchy). Once an ontology term is clicked, the detailed information about the ontology term will be displayed in another web page. The source page is generated using RDF/XML format (Fig. 2B). The second line of the RDF/XML file specifies an XSLT stylesheet in which the HTML code is embedded. The URL of the stylesheet specifies the label abbreviation of a particular ontology (*e.g.*, VO) and an IRI of a specific term (*e.g.*, 'vaccine') in this ontology. Some raw output from SPARQL queries was reformatted by Ontobee to minimize the file size.

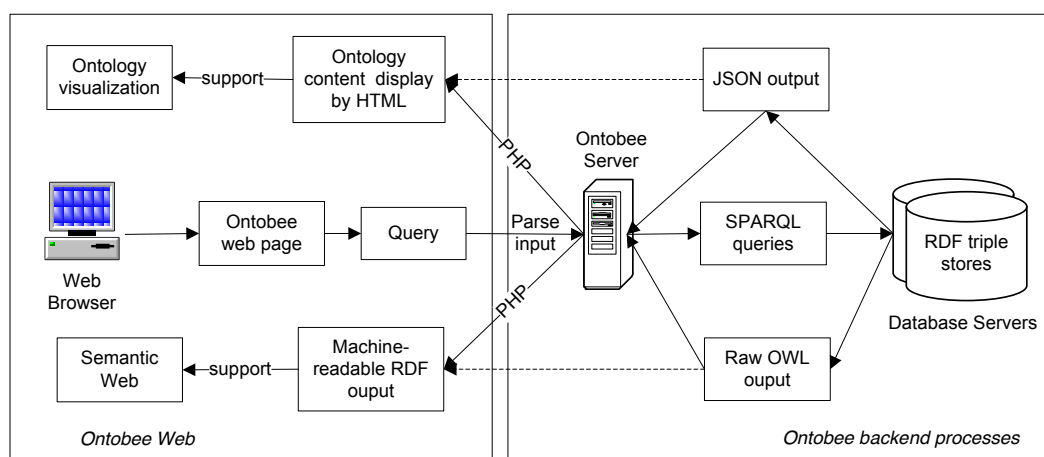


Figure 1. Ontobee architecture design.

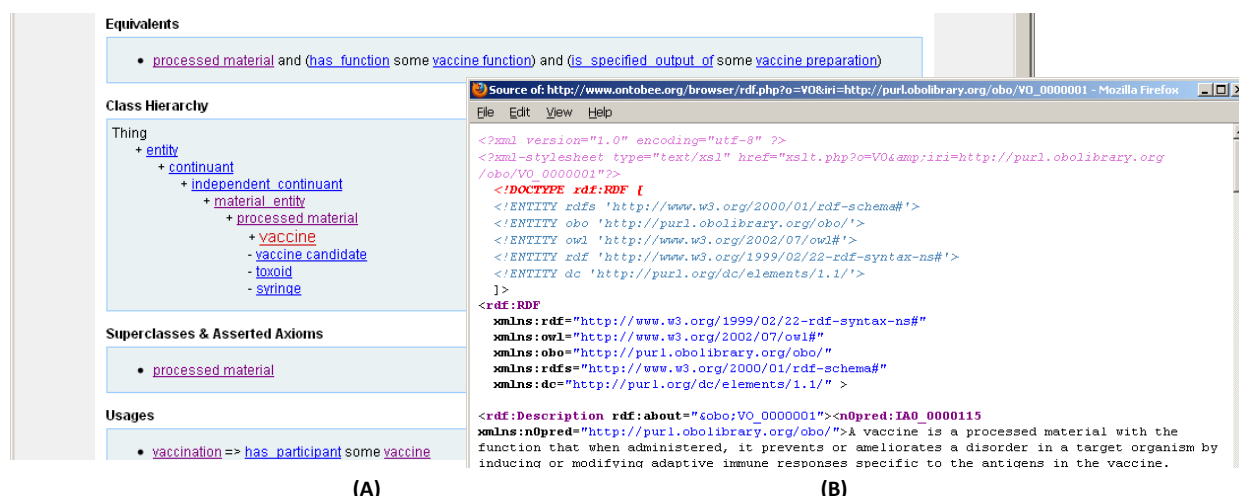


Figure 2. Screenshots of Ontobee display of a VO term 'vaccine' in Firefox web browser. (A) The web HTML display of this term. (B) Source view of this HTML page.

4 Discussion and Summary

The problem of identity and reference of ontology terms on the web has been discussed for long in and outside WWW circles. Ontobee is an attempt to unify biomedical ontology visualization and ontology term transfer based on RDF/XML format. Ontobee provides a useful tool to support the LOD and the Semantic Web.

Acknowledgments

This project is supported by NIH grant 1R01AI081062.

References

1. Cote RG, Jones P, Martens L, Apweiler R, Hermjakob H. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic Acids Res* 2008 Jul 1;36(Web Server issue):W372-6.
2. Ruttenberg A, Rees JA, Samwald M, Marshall MS. Life sciences on the Semantic Web: the Neurocommons and beyond. *Brief Bioinform* 2009 Mar;10(2):193-204.

ICBO Software Demonstrations



ICBO

International Conference on Biomedical Ontology

July 28-30, 2011
Buffalo, New York, USA

Protein-Centric Connection of Biomedical Knowledge: Protein Ontology (PRO) Research and Annotation Tools

Cecilia N. Arighi¹, Darren A. Natale², Judith A. Blake³, Carol J. Bult³, Michael Caudy⁴,
Alexander D. Diehl⁵, Harold J. Drabkin³, Peter D'Eustachio⁶, Alexei Evsikov³,
Hongzhan Huang¹, Natalia V. Roberts¹, Alan Ruttenberg⁷, Barry Smith⁸, Jian Zhang², Cathy H. Wu^{1, 2}

¹Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA

²Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA

³The Jackson Laboratory, Bar Harbor, ME, USA

⁴Ontario Institute for Cancer Research, Toronto, ON, Canada

⁵Department of Neurology, University at Buffalo, NY, USA

⁶New York University School of Medicine, New York, NY, USA

⁷School of Dental Medicine, University at Buffalo, NY, USA

⁸Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, NY, USA

Abstract. The Protein Ontology (PRO) web resource provides an integrative framework for protein-centric exploration and enables specific and precise annotation of proteins and protein complexes based on PRO. Functionalities include: browsing, searching and retrieving, terms, displaying selected terms in OBO or OWL format, and supporting URIs. In addition, the PRO website offers multiple ways for the user to request, submit, or modify terms and/or annotation. We will demonstrate the use of these tools for protein research and annotation.

1 The Protein Ontology Resources

The Protein Ontology (PRO) is a formal and well-principled Open Biomedical Ontologies (OBO) Foundry ontology for proteins and protein complexes [1]. It is one of the first six ontologies recommended by the OBO Foundry as preferred targets for community convergence, alongside the Gene Ontology (GO). The PRO website (<http://pir.georgetown.edu/pro/pro.shtml>) provides an integrative framework for protein-centric exploration and enables specific and precise annotation of proteins and protein complexes based on PRO. The website functionalities include: i) browsing the ontology while displaying selected data, ii) retrieving a specific branch of the ontology, iii) searching the ontology, mappings and annotations, iv) displaying OBO stanzas for selected terms which can be used into visualization tools such as Cytoscape for an integrated view, and v) downloading selected terms in OWL format for import into an ontology or OWL-aware environment. In addition, each term has a corresponding PRO entry report that links the ontology information, the annotations and the mapping to external resources, therefore displaying all

the information available for that term. For example, a term for a given complex will contain relationships and links to all the individual protein components plus annotation that applies to this complex (**Fig. 1**). PRO identifiers are URIs following the OBO Foundry ID Policy (<http://obofoundry.org/id-policy.shtml>). An example is:

http://purl.obolibrary.org/obo/PR_000000000.

URLs are resolvable, providing information in the web browser and linked data access [2] using Ontobee (<http://ontobee.org>).

PRO allows researchers to explore functional and evolutionary relationships of proteins and protein complexes as well as their higher level organization in pathways and protein networks (**Figs. 1 and 2**). For example, **Fig. 2** shows in a single Cytoscape view that glutaminase 1 has a paralog glutaminase 2 (both share the glutaminase domain as shown in annotation of the parent term), that both are found *E.coli* and *B. subtilis*. It also shows the acetylation of glutaminase 1 and that the active glutaminase 1 is a complex (see corresponding annotation) and it is also observed in both species. A controlled vocabulary is used for annotation and PRO interoperates with GO for PRO complexes.

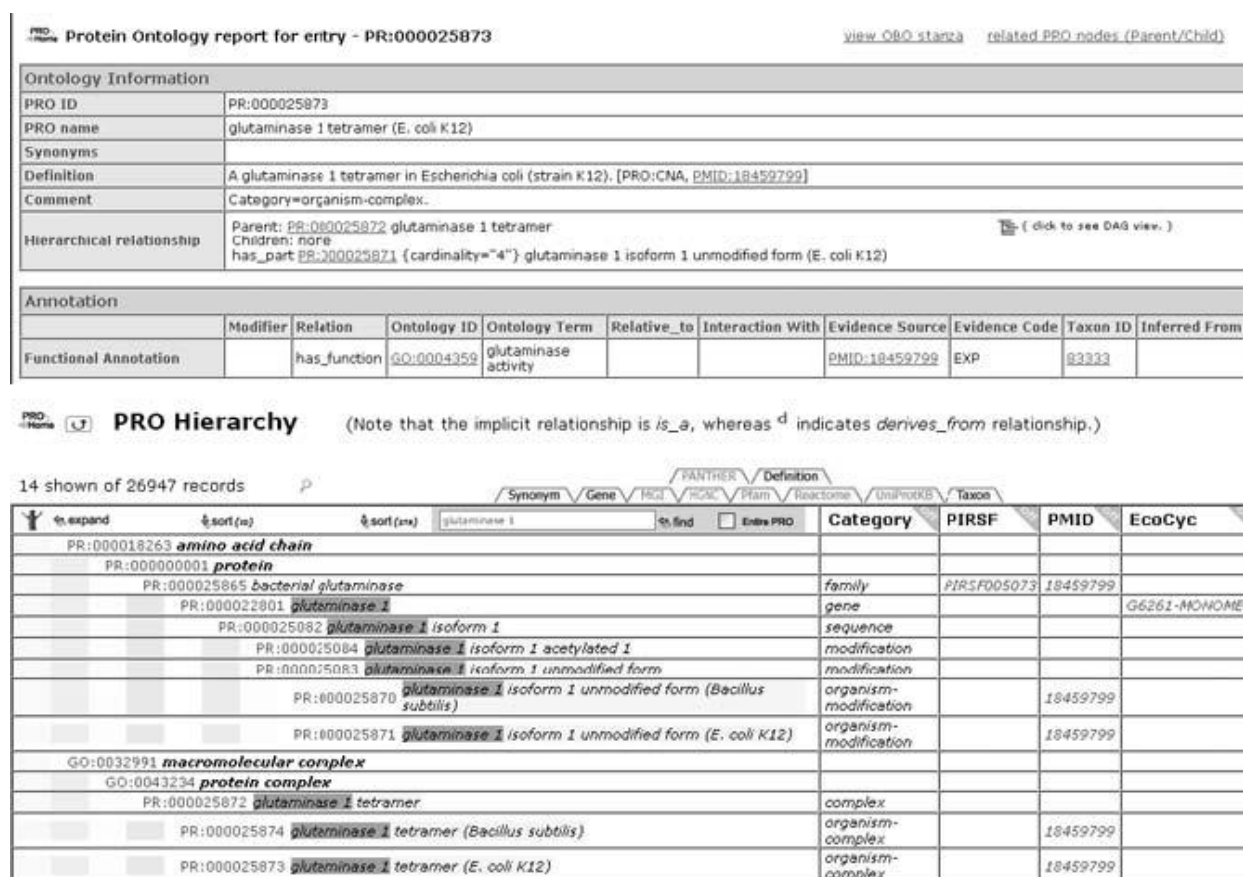


Figure 1. Sample entry report for a glutaminase 1 complex (*upper panel*), the hierarchy of terms related to glutaminase 1 along with selected data (*bottom panel*).

To respond to community needs, the PRO website offers different ways for the user to request, submit or modify terms. A SourceForge tracker can be used to request new PRO terms or modifications of existing ones. Users can submit a request of a few terms or submit a file using a standardized format that can be input into a semi-automatic pipeline for generation of the PRO terms. In addition, domain experts can be actively engaged in the ontology and annotation by submitting to RACE-PRO. This tool allows non-ontologists to author terms and/or annotations. RACE-PRO provides a simple mechanism where the user typically retrieves a protein sequence for the protein form to be described, specifies a sequence region and/or post-translational modification(s) that occurs in the protein form, includes the data source of information (such as PubMed ID), and if need be, adds annotation using controlled

vocabulary (Gene Ontology, MIM, Pfam, Sequence Ontology) (**Fig. 3**). The user is given a reference number to track the annotation and the information is saved internally as a tab delimited file in a format similar to the PRO annotation file (PAF), and checked by a PRO editor. Since most PRO term names and definitions follow a standardized format, a script converts the information therein into PRO terms, by checking for existing terms, and adding parent terms as needed. Once a PRO ID is generated, it is sent to the user along with the PRO term and annotation for a final check and then it is integrated in the PRO release (based on the example in **Fig. 3** two terms were created PR:000026785 and the parent term PR:000002439). We will demonstrate use of these tools to assist protein research and PRO curation.

SHIRAZ and CABERNET: Leveraging Automation, Crowdsourcing, and Ontologies to Improve the Accuracy and Throughput of Zebrafish Histological Phenotype Annotations

Brian Canada¹, Georgia Thomas², John Schleicher³, James Z. Wang³, Keith C. Cheng²

¹ University of South Carolina Beaufort, Bluffton, SC, USA

² Penn State College of Medicine, Hershey, PA, USA

³ The Pennsylvania State University, University Park, PA, USA

Abstract. One of the goals of the Zebrafish Phenome Project is to systematically annotate the cellular-level morphological phenotypes associated with each gene in the zebrafish genome. Here, we offer demonstrations of two complementary software tools designed to help achieve that objective: SHIRAZ, a content-based image retrieval system designed for automated high-throughput annotation of histological phenotypes in the larval zebrafish, and CABERNET, a “crowdsourcing” application for histology image tagging that enables multiple domain experts to achieve consensus on ontology-compliant phenotype annotations. Potentially, such “consensus annotations” not only can be used to improve the accuracy of ground truth data for training SHIRAZ, but they can also be imported directly into PATO-compatible phenomic databases such as the Zebrafish Information Network.

Keywords: content-based image retrieval, crowdsourcing, phenotype annotation, histology, Zebrafish Phenome Project, Phenotype and Trait Ontology, PATO

The Zebrafish Phenome Project [1] aims to produce a comprehensive, ontology-compliant workup of the morphological, behavioral, and physiological phenotypes associated with mutation, environmental, and toxicological effects on each of the 20,000-25,000 genes in the zebrafish genome. Researchers attending the March 2010 Zebrafish Phenome Project community meeting concurred with our (the Cheng lab’s) opinion that annotated high-resolution imaging at the cellular level – such as by histology – was critical for morphological screening, particularly during larval development. While the technology for producing high-resolution histological imaging exists for zebrafish [2], the qualitative aspects of current histological assessments can result in intra- and inter-observer variability caused by differences in training, ability, timing, fatigue, and experience. Consequently, image datasets associated with high-throughput projects such as the Zebrafish Phenome Project will require analysis by *automated* approaches if reproducible cellular-level phenotypes are to be obtained for each gene in a reasonable timeframe.

Recently, we introduced a working prototype of a content-based image retrieval system, called SHIRAZ (System of Histological Image Retrieval and Annotation for Zoomorphology; online demo available at: <http://shiraz.ist.psu.edu>) [3], which has been designed to be capable of automatically annotating high-resolution images depicting histological abnormalities in the larval zebrafish, with a pilot application involving the developing eye at 5dpf. When a “query” eye image is uploaded to SHIRAZ, its automatically-extracted profile of texture features is compared to *feature signatures* associated with a set of ground truth “annotation concepts.” These concepts were derived from our Phenotype and Trait Ontology (PATO)-compliant knowledge base of manually-characterized semi-quantitative phenotype abnormality scores (completed for approximately 100 cloned mutants), ranging from 0 to 4 in order of increasing abnormality [Cheng lab, unpublished data]. SHIRAZ presents its results as a list of predicted annotation concepts, ranked by overall feature similarity to the query image.

While the current SHIRAZ prototype constitutes a viable proof-of-concept for an automated high-throughput histological analysis laboratory workflow, its accuracy is limited by the reliability of the ground truth annotations used in training its classification model. The present model is based on phenotype scores recorded and self-validated by a single observer with an appropriate level of domain knowledge. However, other experts, with varying levels of knowledge and experience, may interpret these same phenotypes differently. If multiple observers can achieve a consensus on the abnormality score of a given phenotype, then one can be more confident in the accuracy of the annotation, thereby improving ground truth data quality. One potential mechanism for achieving this consensus is by *crowdsourcing*, in which large-scale tasks are distributed to communities of human workers. For example, in the popular “ESP Game” for crowdsourced image labeling [4], two randomly paired “players” are presented with an image and, without communicating, must try to guess possible words to label that image. If both players successfully guess the same word, their “game score” increases, and the agreed-upon label is saved in a database for later use. The ESP Game therefore provides a mechanism for recording accurate, non-trivial image annotations in exchange for user entertainment.

To help improve the accuracy of the ground truth annotations used in SHIRAZ, we have proposed our own crowdsourcing application, called CABERNET (Crowdsourcing the Annotation of Bio-images for Education, Retrieval, and Network-Enabled Telephenotyping; current demo at: <http://shiraz.ist.psu.edu/cabernet>). CABERNET uses a controlled interface in which a registered user chooses a zebrafish histology “virtual slide” (powered by the Penn State Zebrafish Atlas [5]) and selects an abnormality score for each ontology-compliant

(e.g., PATO-based) phenotype to be characterized. Each user’s “game score” is dynamically updated based on the degree to which his or her chosen abnormality scores match those selected by other users. Abnormality scores with the highest consensus will be used as ground truth for later re-training of SHIRAZ to help improve its prediction accuracy. Meanwhile, these ontology-compliant “consensus annotations” will potentially form a rich phenotype dataset that, with minimal (if any) modification, can be directly imported into phenomic databases such as ZFIN [6]. Therefore, we expect that phenotypic annotations generated using both SHIRAZ and CABERNET can expedite the progress of the Zebrafish Phenome Project and help to close the genome-phenome knowledge gap.

References

1. Zebrafish Phenome Project 2010 Meeting, <http://www.blsmeetings.net/zebrafish/>
2. Mohideen, M.A., Beckwith, L.G., Tsao-Wu, G.S., Moore, J.A., Wong, A.C., Chinoy, M.R., Cheng, K.C.: Histology-based screen for zebrafish mutants with abnormal cell differentiation. *Dev. Dyn.* 228, 414–423 (2003)
3. Canada, B.A., Thomas, G.K., Cheng, K.C., Wang, J.Z.: SHIRAZ: an automated histology image annotation system for zebrafish phenomics. *Multimed. Tools Appl.* 51, 401–440 (2011)
4. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: *Proc. of SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*, pp. 319–326, ACM, New York (2004)
5. Penn State Zebrafish Atlas, <http://www.zfatlas.psu.edu>
6. ZFIN: The Zebrafish Model Organism Database, <http://www.zfin.org>

Biomedical Ontology Matching Using the AgreementMaker System

Isabel F. Cruz¹, Cosmin Stroe¹, Catia Pesquita², Francisco M. Couto², Valerie Cross³

¹ADVIS Laboratory, University of Illinois at Chicago, USA

²Faculdade de Ciências da Universidade de Lisboa, Portugal

³Computer Science and Software Engineering Department, Miami University, Oxford, OH, USA

Abstract. The AgreementMaker ontology matching system, which has been developed in the ADVIS Laboratory at the University of Illinois at Chicago, has been deployed to dozens of users in a variety of domains. In this demo we concentrate on research advances that make the AgreementMaker system particularly suitable for biomedical applications: (1) An extensible architecture; (2) Automatic combination of the results from matching methods; (3) Integrated matching and evaluation; and (4) Support for external vocabularies. AgreementMaker has recently obtained the best results ever in the OAEI Anatomy Track competition.

Ontology Matching System

AgreementMaker supports a wide variety of *matchers* and manual intervention to correct automatically found mappings or add new ones. An object-oriented architecture is used to define a generic matcher, which defers only a few operations to the concrete matcher extensions [2]. Matchers can be composed in series or parallel (see Figure 1(a)), with the system supporting the *automatic selection* of the weights assigned to the results of each matcher. This capability is integrated with an *evaluation component* that is supported by the user interface (see Figure 1(b)) [3].

Matching Anatomy Ontologies

In 2009, AgreementMaker was ranked a close second in the OAEI Anatomy Track competition [4]. It used as an external vocabulary UMLS. In 2010 it was ranked first and the results that were obtained were the best ever [6]. The improvement over the previous year was partly due to the use of a principled approach for the integration of *vocabularies* [5]. In particular, each matcher was extended to incorporate lexicons. As external vocabulary, only WordNet was used this time.

Collaboration

Several capabilities of the system were driven by the real-world problems of end users who are sophisticated domain experts. To foster this kind of interaction, AgreementMaker is available for download (www.agreementmaker.org). For example, with the University of Lisbon, the *extensibility* of AgreementMaker was tested. They developed new matchers that could be readily incorporated into a new configuration for the Anatomy Track competition [7]. We also gained a better understanding of the nature of the ontologies used in the Anatomy Track (and in particular of their annotations), which led to incorporating lexicons into the matchers. With Miami University, two biomedical ontologies with greater heterogeneity than those of the Anatomy Track were considered. The idea of exploring the nature of the ontologies and of using lexicons was further explored by using annotation profiling [1].

Demo Description

Conference participants can load their own ontologies or use those already mentioned. They can also explore the system freely or follow a walk-through, consisting of browsing the ontologies, running different combinations of matchers, using the annotation profiling component, and evaluating the quality of the matchings.

BioPortal: Ontologies and Integrated Data Resources at the Click of a Mouse

Patricia L. Whetzel¹, Natasha Noy¹, Nigam Shah¹, Paul Alexander¹, Michael Dorf¹, Ray Ferguson¹, Margaret-Anne Storey², Barry Smith³, Chris Chute⁴, Mark Musen¹

¹Stanford Center for Biomedical Informatics Research, Stanford University, CA, USA;

²University of Victoria, Canada; ³University of Buffalo, NY, USA; ⁴Mayo Clinic, MN, USA

BioPortal is a Web portal that provides access to a library of biomedical ontologies and terminologies developed in OWL, RDF(S), OBO format, Protégé frames, and Rich Release Format (<http://bioportal.bioontology.org>). BioPortal functionality, driven by a service-oriented architecture, includes the ability to browse, search and visualize ontologies (Figure 1). The Web interface also facilitates community-based participation in the evaluation and evolution of ontology content. Registered users are able to add mappings between terms, to add comments on individual terms within the ontology, and to provide reviews of ontologies

(Figure 2). This user-generated content provides a critical evaluation and feedback mechanism for ontology developers. BioPortal also enables integrated search of biomedical data resources such as the Gene Expression Omnibus (GEO), ClinicalTrials.gov, and ArrayExpress, through the ontology-based indexing of these resources with ontologies in BioPortal (Figure 3). Thus, BioPortal not only provides investigators, clinicians, and developers a 'one-stop shop' to view and programmatically access biomedical ontologies, but also provides support to integrate data from a variety of biomedical resources.

The screenshot displays the BioPortal Search interface. At the top, there is a search bar with 'melanoma' entered. Below the search bar are options for 'Include attributes in search', 'Contains', and 'Exact Match'. To the right, there are filters for 'Categories' (All Categories), 'Groups' (All Groups), and a 'Filter' text box. Below these filters is a list of ontologies with checkboxes: ABA Adult Mouse Brain (ABA), Adverse Event Ontology (AEO), African Traditional Medicine (ATMO), AIR (AIR), Amino Acid (amino-acid), and Amphibian gross anatomy (AAO). The main section is titled 'Matching Terms' and shows 773 results. It includes a table with columns: Term Name, Term ID, Ontology, Version ID, Ontology ID, Details, and Visualize. The table lists several 'Melanoma' terms from various ontologies like Galen, Cell line ontology, NCI Thesaurus, etc.

Term Name	Term ID	Ontology	Version ID	Ontology ID	Details	Visualize
Melanoma	Melanoma	Galen	4525	1055		
Melanoma	DOID:1909	Cell line ontology	39927	1245		
Melanoma	Melanoma	NCI Thesaurus	42693	1032		
Melanoma	D008545	Descriptores en Ciencias de la Salud (Spanish M...)	42236	1420		
melanoma	11223:68	Health Level Seven	42545	1343		
Melanoma	10053571	MedDRA	42280	1422		
Melanoma	T321	MedlinePlus Health Topics	40397	1347		
Melanoma	C4116	National Drug File	40402	1352		

Figure 1. The BioPortal Search interface. The Search tab allows users to limit their search to “Contains” or “Exact Match” (a) and the ontology content can be limited by “Categories”, “Groups”, or to a specific ontology (b). Search results (c) display the “Term Name”, “Identifier”, and “Ontology Name”. Additional term details are displayed in the term “Details” pop-up and the ontology structure can be viewed in the “Visualize” pop-up.

Notes

Show entries

Filtering Options - ☐ Hide Archived

Search:

SUBJECT	AUTHOR	TYPE	TARGET	CREATED
New Term Proposal: Exercise Study Facility <small>archived</small>	Mette	New Term Proposal	Physiology Facility (Class)	06/15/2010
New Term Proposal: Sleep Study Facility <small>archived</small>	Mette	New Term Proposal		
Deprecate	whetzel	Comment		
Move in hierarchy	whetzel	Comment		
re-activate this term	whetzel	Comment		
Proposal	whetzel	Comment		
Add Education Service as a Synonym	whetzel	Comment		
Synonym proposal	whetzel	Comment		
Add Image to Data_Resource branch	whetzel	Comment		
Clarify action item	whetzel	Comment		
Clarify action item	whetzel	Comment		

Biomedical Resource Ontology Version 3.1

Physiology Facility | Link Here | Subscribe

View Ontology Summary

Jump To: (Go)

- Activity
 - Area of Research
 - Bioinformatics Information Model
 - Biorepositories
 - Core Collection
 - Core Concept
 - Core Concept Scheme
 - Deprecated Activity
 - Deprecated Area of Research
 - Deprecated Resource
- Resource
 - Funding Resource
 - Information Resource
 - Material Resource
 - Assessment Material Resource
 - Biological Supply Resource
 - Biomedical Supply Resource
 - Facility Core
 - Biosafety Level Facility
 - Cell Biology Facility
 - Fabrication Facility
 - Imaging Facility
 - Molecular Biology Facility
 - Physiology Facility
 - Research Animals Facility
 - Tissue Organ Facility
 - Instrument
 - Laboratory Supply Resource
 - Medical Device
 - Reagent Resource
- People Resource
 - Service Resource
 - Software
 - Training Resource

Notes (1) | Mappings (0) | Resource Index

Show entries

Search:

New Term Proposal: Exercise Study Facility

New Term Proposal submitted by Mette 3 months ago on Physiology Facility in Biomedical Resource Ontology

PREFERRED NAME: Exercise Study Facility

PROPOSED ID:

PARENT: http://bioontology.org/ontologies/BiomedicalResourceOntology.owl#Physiology_Facility

REASON FOR CHANGE: Physiology Facility child

STATUS:

CONTACT INFO:

SYNONYMS:

DEFINITION: A facility or core devoted to exercise studies

Responses [hide all](#) | [show all](#)

RE: New Term Proposal: Exercise Study Facility by whetzel 3 months ago

Can you expand on what an exercise study is?

[reply](#)

Exercise study facility definition by Mette 3 months ago

A facility or core that provides services and equipment related to exercise physiology

[reply](#)

Add Reply

Figure 2. The “Notes” page displays a summary of all Notes posted to an ontology and can be sorted by author, type, term name, and date.

Denosumab (preferred name) from: NCI Thesaurus x

[Search](#)

[Clear](#)

To begin, type in text and select a matching term.
Examples: melanoma, lupus, breast cancer, ...
Click search and select a repository.

Ontology filters

- 0 ABRIS GoldMine
- 0 ArrayExpress
- 37 ClinicalTrials.gov
- 0 Database of Genotypes and Phenotypes
- 0 Gene Expression Omnibus Database
- 0 Online Mendelian Inheritance in Man
- 0 PharmGKB (Disease)
- 0 PharmGKB (Gene)
- 38 PubMed
- 0 Research Commons
- 0 UniProt KB
- 0 CalBiochem
- 12 Adverse Event Reporting System Data
- 0 Bioinformatics
- 0 Conserved Domain Database (CDD)
- 0 DrugBank
- 0 HICAD
- 0 Pathway Commons
- 0 PharmGKB (Drug)
- 0 PubChem
- 0 Resarchme
- 0 Bioinformatics
- 0 Bioinformatics
- 0 Bioinformatics

Figure 3. The “All Resources” tab allows users to search for data records tagged with ontology terms of interest and to find related records in other resources via shared ontology annotations.

Populous: A Tool for Populating an Ontology

Simon Jupp¹, Matthew Horridge¹, Luigi Iannone¹, Julie Klein², Stuart Owen¹,
Joost Schanstra², Katy Wolstencroft¹, Robert Stevens¹

¹School of Computer Science, University of Manchester, UK

²Institut National de la Santé et de la Recherche Médicale, Toulouse, France

Abstract. We present Populous, an open source application for gathering content for an ontology and populating that ontology *en masse*. Populous presents authors with a table-based form where columns are tied to take values from particular ontologies; the user can select a concept from an ontology via its meaningful label to give a value for a given entity. Populated tables are fed into templates that can then be used to generate the ontology's axioms. Populous separates knowledge gathering from the conceptualisation; it also removes users from the usual ontology authoring tools.

Availability: Download, source and video via <http://www.e-lico.eu/populous>.

Ontology building environments such as Protégé and OBOEdit offer facilities for the manual authoring of axioms. Such tools are vital for capturing an ontology's form. But there are many ontologies which are very large, with considerable portions formed of repetitions of the same pattern of axioms, varying only in the fillers within that pattern. To avoid the tedium and potential errors of doing this manually, templates can be filled and the axioms for the pattern generated, avoiding the manual authoring of many axioms.

Populous [1] does this by presenting a familiar form-filling table-based user interface for any ontology authors to populate ontology patterns or templates. Rows are tied to the entities being described; columns are tied to properties and the cells constrained to take values from particular ontologies or fragments of an ontology. As an author fills out the template, he or she is guided to place appropriate values within the template. The content of this table can then be transformed into the axioms of the target ontology with an OWL scripting language.

Populous is an extension of RightField¹, which is used for creating Excel documents that contain ontology based restrictions on a spreadsheet's content. RightField is primarily designed for generating spreadsheet templates for data annotation; Populous extends

RightField to support knowledge gathering and ontology generation. Populous and RightField are both open source, cross platform Java applications released under the BSD license. They use the Apache-POI² for interacting with Microsoft documents and manipulating Excel spreadsheets.

Both OWL and OBO ontologies can be uploaded into Populous. Users can also browse and load ontologies directly from BioPortal. Once the ontologies are loaded they are classified by a reasoner and the basic class hierarchy can be inspected. Terms can be selected from the ontology to create validation sets for values that are permitted for a particular selection of cells in the table. Labels from an ontology's entities can be used within a cell, not just URI or URI fragments. Populous allows the addition of free text, even if the cell has an associated validation range; these values are highlighted in red and can act as placeholders for new or suggested terms when no suitable candidate can be found in the validation set.

Populous supports the use of the Ontology Pre-Processor Language³ patterns in order to generate new OWL axioms from the populated template. OPPL is an extension of Manchester OWL Syntax to select, add and remove axioms and it has an interpreter for scripts that

¹ <http://www.rightfield.org.uk>

² <http://poi.apache.org>

³ OPPL, <http://oppl2.sourceforge.net/>

manipulate the ontology. Variables from the OPPL pattern are mapped to columns from the table using the column name through the Populous pattern Wizard.

We have used Populous with biologists to populate large portions of a kidney and urinary pathway ontology [2]. Populous is another piece in the ‘jigsaw’ of tools that support the ontology authoring process. It starts to fill the gap between the term request system and the manual axiom authoring systems by providing a mechanism for ‘filling out’ templates in such a way that they can be validated against the ontologies with which the ontology is being composed. We see Populous as a means for engaging domain experts who are not ontology experts in the authoring process and any ontology author to more effectively populate their ontology’s content.

Acknowledgements

We acknowledge Mikel Egaña Aranguren for his advice, requirements and testing of Populous. This work was funded by the e-LICO project – EU/FP7/ICT-2007.4.4 and by SysMO-DB – BBSRC grant BBG0102181.

References

1. Simon Jupp, Matthew Horridge, Luigi Iannone, Julie Klein, Stuart Owen, Joost Schanstra, Robert Stevens, and Katy Wolstencroft. Populous: A tool for populating ontology templates. In *Semantic Web Applications and tools for the Life Sciences (SWAT4LS)*, Dec 2010.
2. Simon Jupp, Julie Klein, Joost Schanstra, and Robert Stevens. Developing a Kidney and Urinary Pathway Knowledge Base. In *Bio-ontologies SIG*, 2010.

The Vitro Integrated Ontology Editor and Semantic Web Application

Brian Lowe, Brian Caruso, Nick Cappadona, Miles Worthington,
Stella Mitchell, Jon Corson-Rikert, and VIVO Collaboration

Albert R. Mann Library, Cornell University, Ithaca, NY, USA

Abstract. Vitro is an open-source, community-driven semantic web application development platform best known as the software underlying the VIVO researcher networking tool (<http://vivoweb.org>). Vitro has been developed since 2003 primarily to support VIVO, first at Cornell and since 2009 as a scientist networking platform for the NIH-funded VIVO Consortium of seven universities, research institutes, and medical schools in the U.S., VIVO: Enabling National Networking of Scientists. VIVO integrates the Vitro software with the VIVO core ontology and a thin software layer to support editing functions and visual theming specific to that ontology. Vitro provides three major functions in a single web-based tool: OWL ontology creation, import and editing; import or interactive creation and editing of RDF content conformant to the ontology; and display of the content in a public website with navigation, search, and browse features while also serving linked data to semantic web clients. Vitro makes it possible to develop ontologies and populate instance data for public-facing web applications within a single web platform. Because the results of ontology modifications are immediately reflected in the user interface, it also serves as a useful tool for distributed and collaborative cycles of ontology creation, population, review, and revision.

Keywords: VIVO, OWL ontology editor, linked open data, Drupal

Vitro [1] is an open-source, community-driven semantic web application development platform best known as the software underlying the VIVO researcher networking tool (<http://vivoweb.org>). Vitro has been developed since 2003 primarily to support VIVO, first at Cornell and since 2009 as a scientist networking [2] platform for the NIH-funded VIVO Consortium of seven universities, research institutes, and medical schools in the U.S., VIVO: Enabling National Networking of Scientists [3]. VIVO integrates the Vitro software with the VIVO core ontology and a thin software layer to support editing functions and visual theming specific to that ontology.

Vitro provides three major functions in a single web-based tool: OWL ontology creation, import and editing; import or interactive creation and editing of RDF content conformant to the ontology; and display of the content in a public website with navigation, search, and browse features while also serving linked data to semantic web clients. Vitro makes it possible to develop ontologies and populate instance data for public-facing web applications within a single web platform. Because the results of ontology

modifications are immediately reflected in the user interface, it also serves as a useful tool for distributed and collaborative cycles of ontology creation, population, review, and revision.

For applications requiring additional functionality offered by a complete content management platform, the open-source RDFimporter module [4] for Drupal [5] has been developed to pull remote RDF resources from Vitro or other RDF sources and map their content to Drupal nodes. The College of Agriculture and Life Sciences Research and Impact portal [6] at Cornell demonstrates the re-use of Vitro-hosted RDF in Drupal with the addition of map views and customized faceting of search results.

An internal authorization system provides role-based control over core system functions and ontology or content editing actions, and Vitro has been successfully linked to institutional authentication systems using Kerberos and Shibboleth for end-user content editing.

Vitro is a Java application for the Tomcat servlet container. It uses the Pellet library [7] for reasoning and the Jena library [8] to store

the ontology and instance data, and can be configured to use the variety of database backends supported by Jena.

In addition to the VIVO project, Vitro software is being used by the Data Staging Repository (DataStaR) [9] at Cornell and has been adapted and extended by groups in Australia [10] and China [11], [12].

The Vitro open-source platform is available for checkout from Subversion at <http://vivo.sourceforge.net>. A short video demonstration of VIVO is available at <http://vivoweb.org/video-library>.

Acknowledgement

VIVO is supported by grant U24RR029822 from the National Institutes of Health (NIH).

References

1. <http://vitro.mannlib.cornell.edu>
2. Recovery Act 2009 Limited Competition: Enabling National Networking of Scientists and Resource Discovery (U24), <http://grants.nih.gov/grants/guide/rfafiles/RFA-RR-09-009.html>
3. VIVO grant, <http://www.nih.gov/news/health/nov2009/ncrr-02.htm>
4. RDFImporter overview and download page, <https://github.com/milesworthington/rdfimporter>
5. Drupal homepage, <http://drupal.org/>
6. College of Agriculture and Life Sciences Research and Impact portal, <http://impact.cals.cornell.edu>
7. Pellet OWL reasoner for Java, <http://clarkparsia.com/pellet/>
8. Jena – A Semantic Web Framework for Java, <http://jena.sourceforge.net/>
9. Data Staging Repository (DataStaR) at Cornell, <http://datastar.mannlib.cornell.edu>
10. University of Melbourne Research Data Registry, <https://rdr.unimelb.edu.au/vivo/>
11. Subject Knowledge Environment, <http://ske.las.ac.cn/>
12. Biomedical and Health Knowledge Environment, <http://health.las.ac.cn/>

The BioPortal Import Plugin for Protégé

Jithun Nair, Tania Tudorache, Trish Whetzel, Natalya Noy, Mark Musen

Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA

Reusing terms from other ontologies is an essential part of the ontology development process. Ideally, the ability to reuse terms from other sources should be naturally supported in any ontology development environment. The BioPortal Import Plugin is integrated into the Protégé ontology editor¹ and supports the import of classes from ontologies and terminologies stored in BioPortal² – an open repository of over 250 biomedical ontologies and terminologies. The Bio-Portal import, unlike the OWL import, copies a class from the BioPortal source ontology to the local ontology together with the selected set of properties. The source class and the imported (copied) class share the same ID (class IRI), and have hence the same identity.

As projects have different requirements for the import process, we have made the plugin generic and configurable. The main features of the BioPortal Import plugin include:

1. Import only a class or a sub-tree of classes up to the desired depth,
2. Import into the current ontology, or into a new or existing imported ontology,
3. Import the preferred label, synonyms or definitions for a term, and also specify the local annotation properties, if needed,
4. Import metadata for the imported classes or ontologies (e.g., import author, timestamp, BioPortal version, url, and so on),
5. Store the current import configuration for later use in other Protégé working sessions.

Figure 1 shows the basic steps involved in using the plugin to import classes from a BioPortal ontology. The plugin is implemented as a Protégé project plugin and it integrates naturally in the toolbar of the OWL Classes Tab. The plugin uses the BioPortal RESTful services to show a list of all ontologies and their content from BioPortal right in the Protégé user interface (Fig. 1). The user can select one of the ontologies as the source for the import, and a class, which can be imported with a simple button click into the local ontology. The user is also able to customize the import by clicking on the *Configure import ...* button that will bring up the configuration dialog shown in Fig. 2. Once a user makes a configuration, it will be stored as part of the Protégé project file, and can be used for other similar imports. In a future version, we plan to make the import configuration executable, so that the same import can be run again on the same local ontology, for example, if a new version of the source ontology is available in BioPortal.

The BioPortal Import Plugin is open source and available for download from:

http://protegewiki.stanford.edu/wiki/BioPortal_Import_Plugin.

¹ <http://protege.stanford.edu>

² <http://bioportal.bioontology.org>

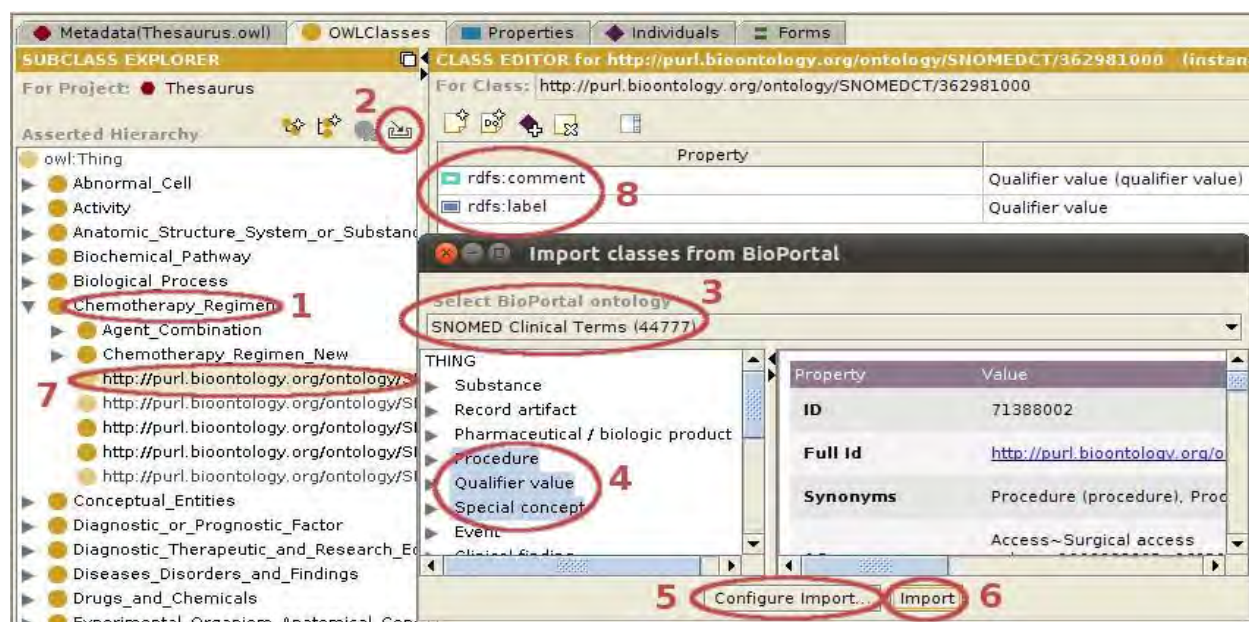


Figure 1. The steps for importing classes from a BioPortal ontology using the BioPortal Import Plugin are as follows: (1) Select a class of a local ontology from the OWLClasses tab. (2) Invoke the plugin by clicking on the icon. (3) Select a BioPortal ontology to import classes from. (4) Browse the BioPortal ontology and select one or more classes to import. (5) If required, change the import settings by clicking on *Configure Import...* (6) Click on *Import*. (7) The imported classes show up as subclasses of the class selected in (1). (8) shows the imported property values for each imported class.

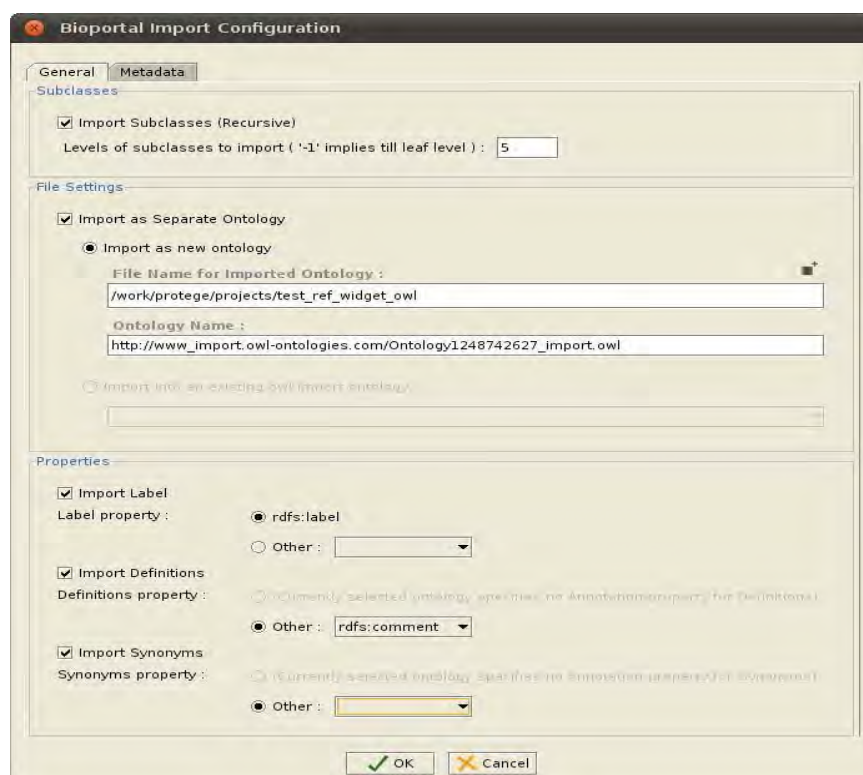


Figure 2. The import configuration for the BioPortal import. Users may configure the import depth, the import target (current ontology, new or existing import) and the properties to import. They may also specify the metadata to import for each imported class or ontology in the Metadata tab.

The Biomedical Ontology Applications (BOA) Framework

Bruno Tavares, Hugo P. Bastos, Daniel Faria, João D. Ferreira,
Tiago Grego, Catia Pesquita, Francisco M. Couto

Faculty of Sciences, University of Lisboa, Portugal

Abstract. The Biomedical Ontology Applications (BOA) framework consists of a set of web applications that aim at an effective exploration of biological information and knowledge discovery using Biomedical Ontologies. This paper presents three BOA web applications: ProteInOn for semantic similarity and protein set characterization based on Gene Ontology (GO); CMPSim for semantic similarity of chemical compounds and metabolic pathways using the Chemical Entities of Biological Interest (ChEBI) ontology; and GRYFUN for the visualization, filtering and analysis of GO functional annotation profiles of a given protein family. The web tools and their documentation, including videos, are freely available from <http://xldb.di.fc.ul.pt/wiki/BOA>

Keywords: Biomedical Ontologies, Semantic Similarity, Web Applications

1 Introduction

The Biomedical Ontology Applications (BOA) framework integrates biomedical ontologies and databases with a number of algorithms and visualization tools, focusing on ontology-based semantic similarity. The BOA web tools, ProteInOn, CMPSim and GRYFUN, are interconnected, and currently support functional analyses that encompass chemical compounds, gene products and metabolic pathways.

2 BOA Web Tools

ProteInOn (Protein Interactions & Ontology) integrates data from the Gene Ontology (GO) [1], GOA, IntAct [2] and UniProt [3] databases. It provides eight GO-based semantic similarity measures [4,5] that can be applied to proteins or GO terms across the three GO aspects. It also supports the identification of the most meaningful functional annotations of a given set of proteins, based on the probability of annotation of each GO term, and the retrieval of interacting proteins. The integration of these features supports complex analysis such as candidate gene identification [6].

CMPSim (Chemical and Metabolic Pathway Similarity) uses the ChEBI (Chemical Entities of Biological Interest) [7] ontology to calculate semantic similarity between chemical compounds and also between pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8] by comparing the sets of ChEBI terms associated with them. The integration of ChEBI-based semantic similarity has been shown to improve existing chemical compound classification systems [9]. CMPSim is also able to find the most meaningful chemical compound classes in a set of pathways, highlighting their common properties.

GRYFUN (GRaph analyZer of FUNctional annotation) supports the visualization of functional profiles of protein sets. These profiles are structured as subgraphs of GO, showing the terms annotated to the protein set and allowing a global view of its functional broadness and specificity. They also provide detailed statistics and relevant metrics by integrating data from GO and UniProt, and support the selection of protein subsets sharing a specific profile. GRYFUN is currently restricted to a dataset of carbohydrate-active enzymes.

Acknowledgments

This work was partially supported by the Fundação para a Ciência e Tecnologia through the Multiannual Funding Programme and grants:

SFRH/BD/29797/2006, SFRH/BD/36015/2007, SFRH/BD/42481/2007, SFRH/BD/48035/2008, and SFRH/BD/69345/2010.

It was also supported by the European Commission through the EPIWORK project under the FP7 (Grant #231807).

References

1. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*. 25, 2010, pp. 25–29.
2. Aranda, B., Achuthan, P., Alam-Faruque, Y., et al.: The IntAct molecular interaction database. *Nucleic Acids Research*. 2010, Vol. 38, pp. D525-D531, (2010).
3. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*. 2010, vol. 38, pp. D142-D148, (2010).
4. Pesquita, C., Faria, D., Bastos, H., Ferreira, A., Falcao, A., Couto, F.: Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9(S5), S4, (2008).
5. Pesquita, C., Faria, D., Falcão, A. O., Lord, P., Couto, F. M.: Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), (2009).
6. Bastos, HP, Tavares B, Pesquita C, Faria D, Couto FM: Application of Gene Ontology to Gene Identification In Silico Tools for Gene Discovery. Springer, (2011).
7. Degtyarenko, K., Matos, P., Ennis, M., et al. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*. (2007).
8. Kanehisa, M., Goto, S., Hattori, M., et al.: From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research*. 2006, Vol. 34, p. D354, (2006).
9. Ferreira, J. D., Couto, F. M.: Semantic Similarity for Automatic Classification of Chemical Compounds. *PLoS Computational Biology*, 6(9), (2010).

The NCBO Annotator: Ontology-Based Annotation as a Web Service

Patricia L. Whetzel, Clement Jonquet, Cherie Youn,
Michael Dorf, Ray Ferguson, Mark Musen, Nigam Shah

Stanford Center for Biomedical Informatics Research, Stanford University, CA, USA

The NCBO Annotator Web Service (http://www.bioontology.org/wiki/index.php/Annotator_Web_service) provides a mechanism to generate ontology-based annotations by tagging textual metadata submitted to the Web service (e.g. text of interest to the user) with ontology terms from BioPortal (<http://bioportal.bioontology.org/>). The Web service can be customized for various use cases through selection of parameters such as the choice of ontologies, stop words, and UMLS Semantic Types (Figure 1). The annotation process generates direct annotations and can also generate expanded annotations based on the *is-a* hierarchy of the ontology and/or mappings

between ontologies (Figures 2, 3). The Annotator Web service is a RESTful Web service and a number of sample clients have been developed (http://www.bioontology.org/wiki/index.php/Annotator_Client_Examples), including an Excel Addin. The ease of access to this annotation functionality makes the task of creating ontology-based annotations accessible for any biomedical researcher. Use cases of the NCBO Annotator include triaging literature to prioritize curation of publications and annotating free text data descriptions from life science databases and clinical records to enhance information retrieval.

Figure 1. The NCBO Annotator Web interface on BioPortal.

Users can select which ontologies to use for annotation (a) and results can be filtered to terms within UMLS Semantic Types (b).

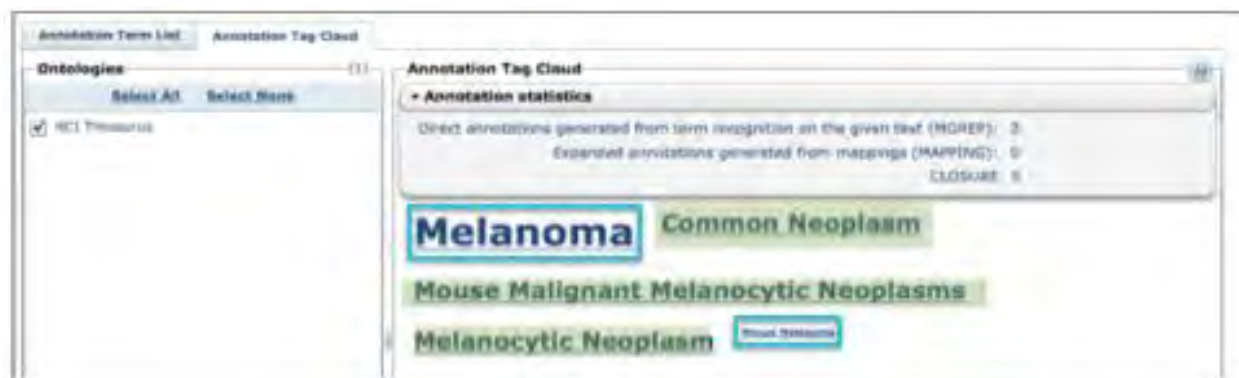


Figure 2. Results from the NCBO Annotator including expanded annotations based on the *is-a* hierarchy of the ontology. The direct annotations are highlighted in blue and the expanded annotations are highlighted in green.

OntoFox and its Application in the Development of the Brucellosis Ontology

Zuoshuang Xiang, Yu Lin, Yongqun He

University of Michigan, Ann Arbor, USA

Abstract. OntoFox (<http://ontofox.hegroup.org/>) is a web-based system to support ontology reuse. OntoFox allows users to input terms, fetch selected properties, annotations, and certain classes of related terms from source ontologies, and save the results using the RDF/XML serialization of the OWL. Currently >80 existing ontologies are available in OntoFox for direct term extraction and importing. OntoFox is an efficient method that promotes ontology sharing and interoperability. As a demonstration, OntoFox has recently been applied for the development of Brucellosis Ontology (BO).

1 Introduction

To avoid duplication of effort during ontology development, it is advised to import pre-existing ontology terms and knowledge into a new ontology if possible. The Web Ontology Language (OWL; <http://www.w3.org/TR/owl-syntax/>) provides a mechanism to import ontologies. This approach leads to import of the whole ontology. However, importing a whole ontology may not be practical and needed, especially when the source ontology is very large and the majority of the terms in the source ontology are not relevant to the new ontology.

To address flexibility of ontology reuse, we have developed OntoFox (<http://ontofox.hegroup.org/>) [1], a web-based application for retrieving ontology term information and importing it into a target ontology. OntoFox follows and expands the principle of MIREOT (Minimum information to reference an external ontology term) [2]. Inspired by existing ontology modularization techniques, OntoFox also develops a new SPARQL-based ontology term extraction algorithm that extracts terms related to a given set of signature terms. In addition, OntoFox provides an option to extract all terms and annotations in an ontology hierarchy rooted at a specified ontology term [1].

As a demonstration, here we show how OntoFox is used to facilitate the development of the Brucellosis Ontology (BO) [3]. BO is a new community-based ontology in the domain

of brucellosis, a zoonotic disease caused by infection of an intracellular Gram-negative bacterium *Brucella* in human and animals. BO is an extension ontology of the Infectious Disease Ontology (IDO) [4]. The development of BO requires a large number of terms from existing ontologies.

2 Features and Usage

The OntoFox input data is processed using PHP, and the processed input data is used in SPARQL queries against an RDF triple store, e.g., the Neurocommons SPARQL endpoint (<http://neurocommons.org>). The returned results are then parsed and converted into an OntoFox output file.

OntoFox provides a user-friendly web form for data input. Alternatively, a user can generate an OntoFox input text file and upload it to OntoFox. A detailed description of OntoFox input file format is available through the OntoFox online tutorial: <http://ontofox.hegroup.org/tutorial>). Briefly, the steps using OntoFox web form to generate OntoFox input include:

1. **Select source ontology.** A dropdown menu can be used to select one ontology out of the available ontologies in OntoFox. If your favorable ontology is not available in OntoFox, a SPARQL endpoint and a graph URI points to the ontology can be specified instead.



Figure 1. OntoFox retrieval of ontology information of 'polymerase chain reaction' from OBI.

(A) A screenshot of how to provide OntoFox input data;

(B) OntoFox input text file that matches the input in (A) and can be automatically generated by the server.

2. **Class term specification.** This can be done in two formats:

a) Bottom up term specification. To implement this, three sets of information are needed: Low level source term URIs, top level source term URIs and target direct superclass URIs, and setting for retrieving intermediate source terms [1].

b) Top-down branch term specification. This feature is designed to extract all terms of an ontology hierarchy branch under a specific ontology term.

3. **Annotation/Axiom Specification.** This allows a user to specify which annotations and axioms to be included in information retrieval.

4. **Specification of the URI of the final OWL (RDF/XML) output file.** The specified URI will be automatically added to the OntoFox output file. If the target ontology includes the same URI information, no additional edition is required for the import of the OntoFox output results into the target ontology.

As an example, we show how to extract the term 'polymerase chain reaction' from OBI into BO (Fig. 1). BO requires the term 'polymerase chain reaction', its relevant

restrictions, and its direct superclass term and top class terms from OBI. After a user types the first few characters of a term, OntoFox automatically provides a list of matching terms for selection, which facilitates the term inputting process (Fig. 1A). After all requirements are specified, an OntoFox input text file will be generated by OntoFox (Fig. 1B).

Under the platform of the Protégé ontology editor (<http://protege.stanford.edu/>), an OntoFox output OWL file can be directly visualized and imported in the target ontology (e.g., BO) using the OWL import function. This approach allows efficient import of minimum information of external ontology terms into target ontology.

In the past year, OntoFox has been used over 2,000 times by more than 700 unique users from 51 countries. We have received many requests and questions during the time. Our software demonstration will detail how the newly developed OntoFox program works and answer some common questions raised by OntoFox users.

3 Use Case: Application of OntoFox in BO Development

Currently BO imports 232 individual terms from eight existing ontologies using OntoFox.

These eight ontologies include: Chemical Entities of Biological Interest (ChEBI), Gene Ontology (GO), Information Artifact Ontology (IAO), NCBI_Taxon, OBI, Ontology for General Medical Science (OGMS), Protein Ontology, and Vaccine Ontology. All OntoFox input files for these ontologies exist on the BO Sourceforge website:

<http://bo-ontology.svn.sourceforge.net/viewvc/bo-ontology/trunk/src/ontology/imports/>.

It is noted that the input files for GO, IAO, and OGMS are not listed in this website. It is because those terms partially imported to BO come from IDO, which is fully imported to BO as its immediate upper level ontology. The ontology term 'polymerase chain reaction' (Fig. 1) is one of the OBI terms imported to BO using OntoFox. The OntoFox output OWL files are located in the parent directory of the above folder.

It is extremely difficult to partially import these terms to BO without such a tool as OntoFox. To our knowledge, the only other tool for such task is MIREOT [2]. However, the usage of MIREOT requires command line programming and does not include a user-friendly interface.

4 Summary

OntoFox is a new web server that automatically extracts external ontology terms and their annotations for efficient ontology development. OntoFox is easy to use without a prior knowledge of SPARQL query and command line programming. OntoFox provides an efficient approach to facilitate ontology sharing and interoperability.

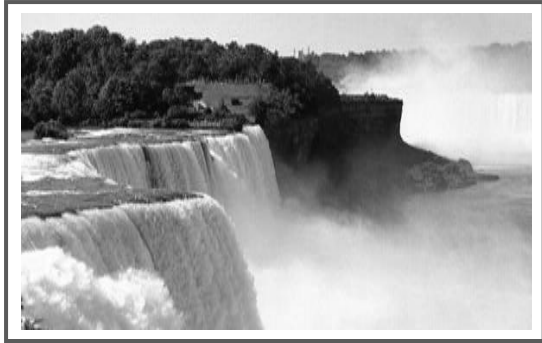
Acknowledgements

This research is supported by a NIH R01 grant (1R01AI081062).

References

1. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y. OntoFox: web-based support for ontology reuse. *BMC Research Notes*. 2010, 3:175.
2. Courtot M, Gibson F, Lister AL, Malone J, Schober D, Brinkman RR, Ruttenberg A. 2011. MIREOT: the Minimum Information to Reference an External Ontology Term. *Applied Ontology*. 6 (1): 23-33.
3. Brucellosis Ontology (BO). Available at: <http://sourceforge.net/projects/bo-ontology>.
4. Infectious Disease Ontology (IDO). Available at: <http://infectiousdiseaseontology.org>.

Representing Adverse Events



ICBO

International Conference on Biomedical Ontology

July 26, 2011
Buffalo, New York, USA

AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events

Yongqun He¹, Zuoshuang Xiang¹, Sirarat Sarntivijai¹, Luca Toldo², Werner Ceusters³

¹University of Michigan Medical School, Ann Arbor, MI, USA

²Merck KGaA, Darmstadt, Germany

³State University of New York at Buffalo, NY, USA

Abstract. The Adverse Event Ontology (AEO) is a realism-based biomedical ontology for adverse events. Currently AEO has 484 representational units annotated by means of terms including 369 AEO-specific terms and 115 terms from existing feeder-ontologies. In AEO, the term ‘adverse event’ is used exclusively to denote pathological bodily processes that are induced by a medical intervention. This requirement for a causal association between an adverse event and a medical intervention clearly distinguishes our approach from other approaches according to which any untoward phenomenon observed to have appeared in a mere temporal relation with some medical intervention becomes reported as an ‘adverse event’. We label such phenomena as being the subject of ‘adverse event hypotheses’.

Keywords: Adverse event, AEO, Adverse Event Ontology

1 Introduction

While medical interventions such as drug and nutritional product administrations, vaccinations, and use of medical devices are applied with the goal of producing positive effects, they might induce unwanted reactions which are typically described as ‘adverse events’ or ‘side effects’. An ideal medical intervention should have high efficacy and no unwanted reactions. It is however well known that any substance (even water) might give rise to unwanted reactions, if administered at the wrong dose.

Adverse event related morbidity and mortality are a major public health issue. To better organize adverse event information, different sorts of systems such as COSTAR, MedDRA, the Common Terminology Criteria for Adverse Events (CTCAE), and the WHO’s Adverse Reaction Terminology (WHO-ART) have been developed many years ago. These systems are typically constructed as controlled vocabularies, terminologies or classification systems. These older systems differ from various newer sorts of artifacts that are known as ‘biomedical ontologies’ and which in most cases are consensus-based controlled vocabularies of terms and relations with associated definitions, which are logically

formulated to promote automated reasoning. Bosquet *et al.*, for instance, have shown that terminological reasoning improves the performance of both data mining [1] and data access [2] in pharmacovigilance databases, and have done preliminary work toward the proposal of a categorial structure for adverse drug reactions (ADRs) [3]. However, although logically formulated definitions and axioms have the capacity to produce *valid* reasoning in deductive logic-based reasoning systems, they do not guarantee *sound* reasoning. Typical for prevailing paradigms in biomedical ontology design is concept-orientation which lacks a formal method to relate representational units to that in reality what they are representations of, and these representations are therefore more vulnerable for mistakes that lead to unsound reasoning [4]. Specifically in the context of what is called ‘adverse event’, there is much diversity in what is considered to be good terminological practice [5] and appropriate ontological analysis with the result that a variety of entities of totally different sorts with labels such as ‘reaction’, ‘effect’, ‘event’, ‘problem’, ‘experience’, ‘injury’, ‘symptom’, ‘illness’, ‘occurrence’, ‘change’, ‘act’, and even ‘something’, ‘observation’ and ‘term’, have been proposed as super-ordinate terms for ‘adverse event’ [6].

The Adverse Event Ontology (AEO), in contrast, is an ongoing realism-based effort that aims to reduce the confusion in adverse event terminology and representation using the framework offered by the OBO Foundry [7]. In this report, we present our current development of AEO, thereby distinguishing it in particular from another recent effort to generate an Adverse Event Reporting Ontology (AERO).

2 Methods

The development of AEO follows the OBO Foundry principles such as openness, collaboration, and use of a common shared syntax [7] in addition to the principles of Ontological Realism [8]. AEO is thus aligned with the Basic Formal Ontology (BFO) [9] and the Relation Ontology (RO) [10].

The AEO development method follows many guidelines provided by Ceusters *et al.* [6] in generating ontological representations of adverse events on the basis of inspecting the sorts of particulars that are involved when something that might be labeled as ‘adverse event’ comes into existence in some patient. These particulars are:

- (1) #1: a medical intervention (*e.g.*, vaccination, drug administration)
- (2) #2: a patient
- (3) t1: the time at which the medical intervention is given to the patient
- (4) #3: a clinically abnormal process (*e.g.*, a fever process)
- (5) t2: the time at which the clinically abnormal process happens

These elements can be modeled in the adverse event design pattern of Fig. 1 which restricts the term ‘adverse event’ to those pathological bodily processes that are induced by a medical intervention.¹ Both *adverse event* and *medical intervention* are subclasses of *processual_entity* as defined in BFO. Instances

of these two processes occur each at a specific temporal region. The corresponding causal relation between the referents of these two process terms is represented using the term *induced_by* in AEO. Such a relation term is not available in RO or any other ontologies. It is noted that the OBI term *process is result of* (OBI_1110060) is for direct causality and not indirect causality as required here. Fig. 1 introduces the basic adverse event at the class level. In clinical cases, instance level modeling can be generated. For example, a specific vaccination process carried out on a particular patient is an *instance_of* a medical intervention. To illustrate this and other important points, an example is provided in the next section. It is also to be noted that the time at which a medical intervention is given to a patient is always earlier than the time at which an adverse event occurs, *i.e.*, t1 *earlier* t2 (this can be made more precise in the context of some guideline, *e.g.*, t1 less-than-4-days-earlier-than t2).

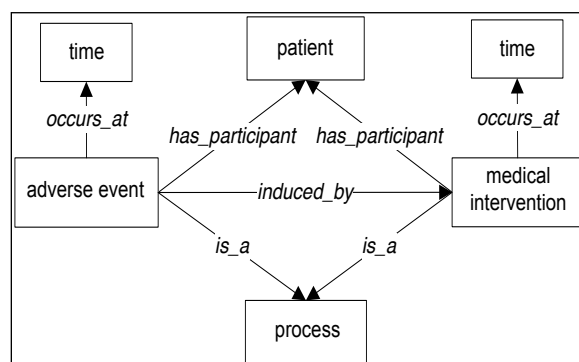


Figure 1. Basic AEO adverse event design pattern.

An OWL version of AEO is developed using Protégé 4. OntoFox [11] was used to extract terms from external ontologies and import them into AEO. For adverse event-specific terms, new identifiers, unique to AEO, were generated.

The latest AEO, although not completely curated in terms of the principles mentioned earlier, is available for public view and download at <http://sourceforge.net/projects/aeo/>. AEO has been submitted to the NCBO BioPortal for public visualization and querying. It is to be noted, however, that this version is a simplification brought about by the fact that OWL, and specifically OWL-DL, does not allow representing that continuants, in contrast to occurrents, exhibit relations in which time is

¹ Although we believe that this more specific meaning of ‘adverse event’ as used within AEO better captures what the entities denoted by this term objectively are and that it would be beneficial that this usage would be generally adopted, the goal of this communication is not to force such usage on the community.

one of the relata, and as a consequence is therefore inadequate for representations that follow these principles.

3 Results

3.1 AEO Statistics

Currently AEO has 484 representational units, annotated by means of 369 terms with specific AEO identifiers, and 115 terms imported from existing ontologies (Table 1). This ontology development design avoids regeneration of new ontology terms that are not in the scope of the adverse event domain and supports efficient ontology reuse on the condition that the feeder ontologies are based on the same principles.

Ontology Names	Classes	Object properties	Total
AEO (Adverse Event Ontology)	368	1	369
BFO (Basic Formal Ontology)	39	0	39
RO (Relation Ontology)	6	25	31
IAO (Information Artifact Ontology)	2	0	2
OBI (Ontology for Biomedical Investigations)	8	3	11
OGMS (Ontology for General Medical Science)	5	0	5
VO (Vaccine Ontology)	19	3	22
NCBITaxon (NCBI Taxonomy)	5	0	5
Total	452	32	484

Table 1. Summary of ontology terms in AEO or imported from existing ontologies.

Existing ontologies are used in two different ways in AEO: one is to import the whole ontology (here BFO and RO), and the other is to import individual terms from existing ontologies. The OntoFox method is a newly developed approach to make individual term importing easy and standardized [11], although additional steps are required to make sure that the definitions for these terms in the feeder-ontologies correspond to the intended referents in AEO.

Fig. 2 lists key terms in AEO. Based on the *adverse event* definition, AEO required the term *medical intervention*, which currently includes four subclasses: *vaccination* (imported

from VO), *drug administration*, *medical device usage*, and *nutritional product usage*. Each of these medical interventions can induce corresponding adverse events, e.g., *vaccine adverse event*.

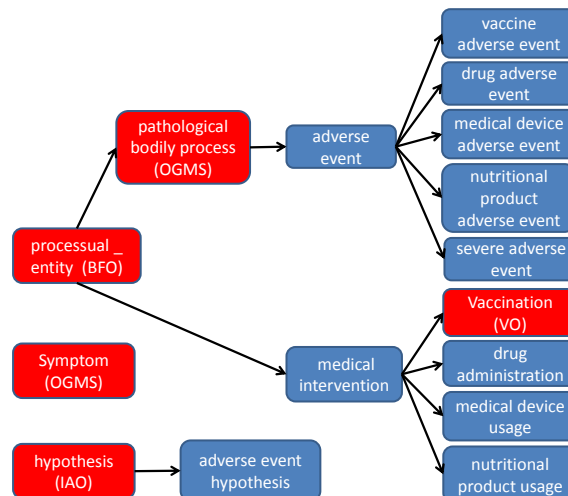


Figure 2. Key representational units in AEO. Dark (red) boxes contain imported terms; Light (blue) boxes are AEO-specific terms.

Instances of adverse event can have outcomes of different types, for instance a sign (e.g., fever, rash) as defined in the Ontology of General Medical Science (OGMS) or another process (e.g., bacterial infection). AEO uses sign- and symptom-related terms (e.g., fever generation) from other existing ontologies such as the Gene Ontology (GO).

3.2 Logical Definition of ‘Adverse Event’ in AEO

The term ‘adverse event’ may mean different things in different settings [6]. In AEO, the term ‘adverse event’ is reserved for those pathological bodily processes that are induced by a medical intervention. As defined in OGMS, a pathological bodily process (OGMS_0000060) is a bodily process that is clinically abnormal. This definition fits well with adverse event and thus is chosen as the parent term of *adverse event* in AEO.

The word ‘induced’ in the AEO ‘adverse event’ definition indicates the existence of a causal chain. A medical intervention is a process in which several independent continuants (e.g., anatomical parts of human body) participate in a variety of ways and of which other processes are parts in which these

or other independent continuants participate. Some independent continuants existed already before the intervention started (*e.g.*, cells and molecules of the patient), others are created (*e.g.*, molecular complexes formed by bodily molecules and drugs) or modified (*e.g.*, opening and closing of membrane channels, folding of proteins) through processes that are part of the intervention or bodily processes that come into existence in response to the creation or modification of these continuants. After the intervention, there are still bodily processes going on in which at least one of the independent continuants just mentioned participates and further independent continuants are created. The term ‘induced’ means that there is at least one chain of processes that starts with some process that is part of the intervention and ends with a pathological bodily process, the chain being further such that for each process within it (except the first one) there is at least one independent continuant that participated or was created in the process immediately preceding it. Note that we are not saying that there is one such independent continuant that participates in the entire chain, but rather something like this:

P1: C1, C2, C3
P2: C2, C4, C5
P3: C5, C6, ...

Mere temporal precedence is not enough because that would allow for chains of processes in which there is a pair that does not ‘share’ at least one continuant.

An alternative definition for ‘adverse event’ would be to assign it as a child term of *ogms:sign*, which has the textual definition of “A quality of a patient, a material entity that is part of a patient, or a processual entity that a patient participates in, any one of which is observed in a physical examination and is deemed by the clinician to be of clinical significance.” Although this appears to cover different adverse events, this *ogms:sign* definition is too broad since all adverse events are processes. At the same time, it is too narrow because there are adverse events that are not observed. The definition of sign in OGMS clearly states “is observed in a physical examination”, instead of “CAN BE observed”.

4 Discussion

Several adverse event representation systems have been proposed thus far while others are under development. For example, the EU-funded project ‘Patient Safety through Intelligent Procedures in medication’ (PSIP) aims to develop innovative tools for generating and providing relevant knowledge to healthcare professionals and patients for ADE prevention. Another relevant project funded by EU is the European Public Warning System (EU-ALERT). The French *VigiTermes* project is an application that automates potential adverse event detection by identifying statistical and semantic links between drugs, treatments and induced pathologies or symptoms. The EU funded *ReMINE* project uses an adverse event ontology to manage patient safety risks in hospital settings [12]. These projects focus on using ontologies in order to facilitate identification of drug related adverse events, combining ontologies with information extraction and also applying ontologies to hospital data.

However, as shown in [6], there is a wide variation in opinions about what would count as adverse event and many definitions fall short in various aspects. Edwards et al for instance define an adverse drug reaction as ‘An appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product’ [3, 13]. The problem with this definition is that it is not specified, for instance, for whom the reaction is unpleasant (appreciation can be different for the patient, his caregivers and his relatives) and that it is prone to, so we assume, unwanted interpretations. Imagine a patient that took an oral overdose of some medicinal product and therefore is subjected to gastric suction to remove what is left in the stomach. Due to erroneous manipulation of the suction device, the patient develops a gastric bleeding. Clearly, this intervention is related to the use of a medicinal product, but it would be wrong to state, although in line with Edwards’ definition, that this gastric bleeding is an adverse drug reaction.

4.1 Adverse Events versus Adverse Event Hypotheses

AEO's requirement of a causal relation between an adverse event and a medical intervention is an important and novel point which removes a lot of ambiguity. The causal requirement is indeed the major aspect in which AEO differs from that of the concept of adverse event as used in existing adverse event reporting systems such as the Vaccine Adverse Event Reporting System (VAERS) and the new Adverse Event Ontology Reporting Ontology (AERO). The latter systems do not require a causal relation to be established between a reported side effect and a medical intervention. Since what is reported as 'adverse event' in these systems may not be truly induced by a medical intervention these adverse event reporting systems contain rather references to pathological processes that happened in a specific timeframe after a medical intervention, some of which might be indeed adverse events in the AEO sense.

The data stored in such an adverse event reporting system is typically used to generate hypotheses about whether what is reported as adverse events and medical interventions are causally linked. Such a hypothesis, represented by the term *adverse event hypothesis* in AEO, becomes critical when a dramatically large amount of cases are reported following the same medical intervention. Therefore, adverse event reporting is not an end. To find potential safety problems is an ultimate goal of reporting adverse events. This is one reason why AEO aims to represent not only the adverse event hypothesis, but also the final causal association.

Finally, note that when a clinician or a patient reports an event after some medical intervention for which it is only later proven to have caused the event, this event does not 'become' an AEO adverse event: it was an instance thereof from the very beginning, although unknown as such until the proof was delivered.

It is possible to reconcile AEO and AERO in a future time. While the events included in AERO for a specific medical intervention may be larger in number than the true adverse events caused by this intervention, AEO has more depth and targets for representation of a knowledgebase of adverse events truly caused

by medical interventions. How to find out the cause-and-effect relation from the reported adverse events in adverse event reporting systems is often a challenge. Rehan *et al.*, for instance, provides physicians with a guide how to assess causal relations of adverse events induced by drug administration [5]. It will surely benefit the public health and has been a critical research topic ever since such an adverse event reporting system is invented.

4.2 Comparison with Other Adverse Event Representation Systems

Here we particularly compare our AEO approach with the representation model for adverse drug reactions (ADRs) provided by Bosquet *et al.* [3].

Bosquet *et al.* generated an ADR model that contains 19 semantic categories, and the categorical structure consists of 8 semantic categories within that model. Sixteen semantic links are described in their ontology. The set of minimal constraints are 4: an ADR should be classified as a disorder, an accident, an investigation, or a syndrome. A structural disorder is defined by at least one location and one morphology. A functional disorder is defined by at least one abnormal function. There are at least one semantic link *is_related_to* and one semantic category "Drug".

The work by Bosquet *et al.* largely differs from ours. First, their ontology is based on categorial design, while AEO is based on OBO foundry ontology design. Second, their approach does not model time dependency between a drug administration and an adverse event. Third, a causal relation between a drug administration and an adverse event is not clearly specified in their system, although it can be assumed to be the case under some interpretation of 'resulting from' in their definition.

4.3 Example: Vaccine-Induced Adverse Events

In the USA, more than 10 million vaccines per year are administrated to children less than 1 year old, usually between 2 and 6 months of age. At this age, infants are at greatest risk for many medical adverse events such as high fevers, seizures, and sudden infant death syndrome.

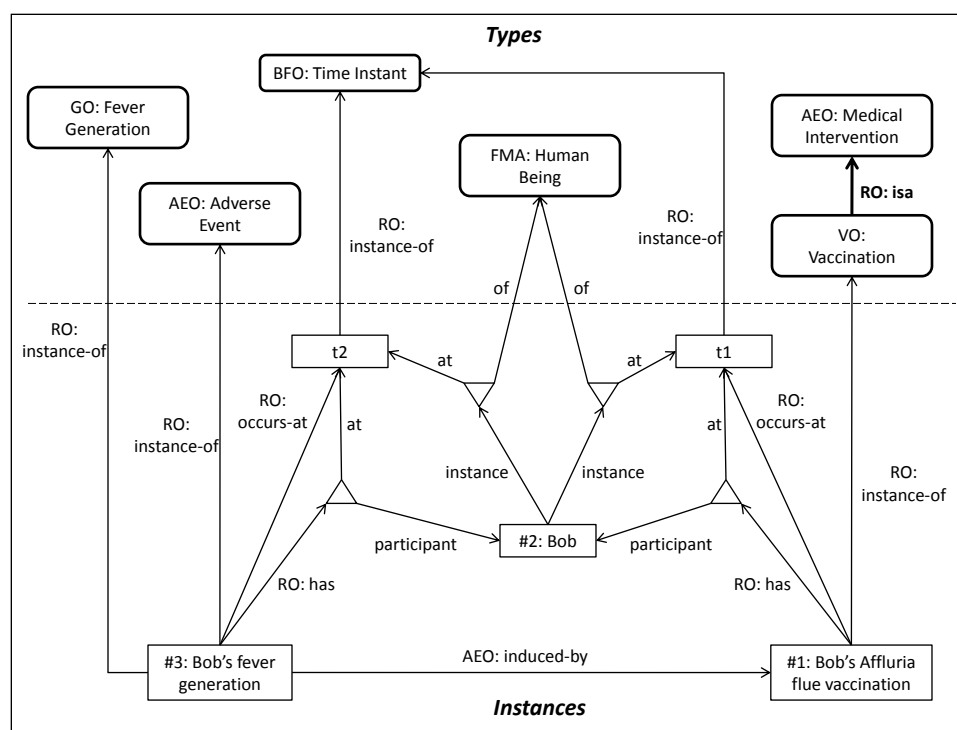


Figure 3. Modeling of vaccination-induced fever adverse event in AEO.

Fig. 3 provides an example of how AEO represents at instance level a specific vaccine-induced fever adverse event. In this example, it is represented that Bob was vaccinated with an Afluria flu vaccination at time t1, and then had a fever at time t2, t1 and t2 being instances of temporal instant. That Bob is an instance of human being at each of these time instants and that he participates at these time instants in the respective processes is represented as well (the little triangles in Fig. 3 indicate that the participation and instance relations involving continuants are three-place relations). Since it is notified in the vaccine instruction that fever generation is an expected adverse event and Bob was in good health before the vaccination, Bob's fever generation is considered as an adverse event induced by the vaccination process. The term *fever generation* is imported from the Gene Ontology (GO).

The Brighton Collaboration is a global research network that set vaccine safety research standards and does not either assume a cause-and-effect relation. According to the Brighton Collaboration, fever is defined as an elevation of body temperature above the normal [14]. Similar to other Brighton

Collaboration definitions, the fever definition itself defines a clinical entity without inference of a causal relation to a given exposure. Therefore, the time interval from immunization until onset of the event cannot be part of the definition itself [14]. However, since AEO assumes such a cause-and-effect relation, this time interval is an important study topic in the AEO representation of an influenza vaccination and a fever vaccine adverse event. Therefore, we argue that AEO and those domain-specific adverse event ontologies aligned with AEO represent a knowledgebase of adverse events caused by medical interventions, where the data stored in regular adverse event reporting systems contain many random (coincident) and false positive events that are not induced by medical interventions.

5 Conclusion

Adverse events endanger patient safety and result in considerable extra healthcare costs. A community-based ontological representation of adverse events is crucial for improving adverse event research. The advent of AEO provides an opportunity for the adverse event

research community to work together towards realism-based adverse event information representation and data analysis.

To monitor and study these adverse events, many vaccine and drug adverse event reporting systems have been established to collect information about adverse events that occur after the administration of licensed vaccines. The examples of national vaccine safety surveillance programs include the VAERS in the USA and the Adverse Events Following Immunization Reporting program by the Public Health Canada. These systems contain reported data about both coincidental events and those truly caused by vaccines. In our view, an ontological representation using AEO will provide a unified and machine-readable representation of various adverse events and support more advanced adverse event data analysis.

Many efforts are required to improve AEO. For example, for better adverse event data representation and knowledgebase establishment it is important to link AEO to adverse event terminologies such as MedDRA and WHO-ART, although caution is here required because of the lack of formal rigor in these systems [15]. It will be challenging and rewarding to predict and identify which events that are temporally associated with medical interventions exhibit causal relations with these interventions using informatics approaches (e.g., statistical algorithms, and literature mining). The drug adverse events are often affected by the genetic background (e.g., SNPs) of the patient. The intricate drug-patient and drug-drug interactions are crucial to determine the final adverse event outcomes. Some adverse events happen due to cross-interactions between drug and non-drugs (e.g., grapefruit). Sometimes, an adverse event emerges when a drug is removed. It would be ideal to model these interactions in AEO with a purpose to understand the fundamental adverse event mechanisms.

Acknowledgments

Development of AEO is supported by NIH grant 1R01AI081062.

References

1. Bousquet, C., *et al.*, *Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach*. Int J Med Inform, 2005. **74**(7-8): p. 563-71.
2. Alecu, I., *et al.*, *PharmARTS: terminology web services for drug safety data coding and retrieval*. Stud Health Technol Inform, 2007. **129**(Pt 1): p. 699-704.
3. Bousquet, C., *et al.*, *Semantic categories and relations for modelling adverse drug reactions towards a categorial structure for pharmacovigilance*. AMIA Annu Symp Proc, 2008: p. 61-5.
4. Klein, G.O. and B. Smith, *Concept Systems and Ontologies: Recommendations for Basic Terminology*. Transactions of the Japanese Society for Artificial Intelligence, 2010. **25**: p. 433-441.
5. Rehan, H.S., D. Chopra, and A.K. Kakkar, *Physician's guide to pharmacovigilance: terminology and causality assessment*. Eur J Intern Med, 2009. **20**(1): p. 3-8.
6. Ceusters, W., *et al.*, *An evolutionary approach to the representation of adverse events*. Stud Health Technol Inform, 2009. **150**: p. 537-41.
7. Smith, B., *et al.*, *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration*. Nat Biotechnol, 2007. **25**(11): p. 1251-5.
8. Smith, B. and W. Ceusters, *Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies*. Applied Ontology, 2010. **5**(3-4): p. 139-188.
9. Bittner, T., Smith, B., *Normalizing Medical Ontologies Using Basic Formal Ontology*. Tagungsband der 49. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, 2004.
10. Smith, B., *et al.*, *Relations in biomedical ontologies*. Genome Biology, 2005. **6**(5): p. R46.
11. Xiang, Z., *et al.*, *OntoFox: web-based support for ontology reuse*. BMC Res Notes, 2010. **3**: p. 175.
12. Ceusters, W., *et al.*, *An Evolutionary Approach to Realism-Based Adverse Event Representations*. Methods of Information in Medicine, 2011. **50**(1): p. 62-73.
13. Edwards, I.R. and J.K. Aronson, *Adverse drug reactions: definitions, diagnosis, and management*. Lancet, 2000. **356**(9237): p. 1255-9.
14. Michael Marcy, S., *et al.*, *Fever as an adverse event following immunization: case definition and guidelines of data collection, analysis, and presentation*. Vaccine, 2004. **22**(5-6): p. 551-6.
15. Bousquet, C., *et al.*, *Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reactions*. Drug Safety, 2005. **28**(1): p. 19-34.

Reporting Adverse Events: Basis for a Common Representation

Mélanie Courtot¹, Ryan R. Brinkman^{1,2}, Alan Ruttenberg³

¹BC Cancer Agency, Vancouver, BC, Canada

²Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

³School of Dental Medicine, University at Buffalo, NY, USA

Abstract. The process of adverse event reporting aims to monitor the status of patients in clinical trials or to provide ongoing monitoring of the safety of interventions once they are in the market. Such reports help identify issues with treatment safety and efficacy, and allow for better education of health practitioners and the general public, ultimately allowing us to learn from our mistakes. However if such reports are to be maximally useful, the information they contain must be unambiguously shared, via standardization and accurate documentation. Towards this end, we briefly review existing reporting standards and then define *adverse event*, as well as other relevant terms, in a manner consistent with the use of the term in existing reporting guidelines. Novel aspects of this work include attention to the distinction between the classification of adverse events based on reporting versus the pathological process types they attempt to monitor, and integration with relevant OBO ontologies to minimize redundant definitional work as well as enable integration of adverse event reporting into the broader landscape of representation for translational medicine. Implementation of a prototype that incorporates this approach is discussed - the Adverse Events Reporting Ontology (AERO).

1 Introduction

In the clinical community, adverse events denote any untoward medical occurrence following a medical intervention. Adverse event reporting is a major part of clinical research, and an important tool to improve patient safety. By collecting and analyzing adverse events we can better understand and prevent them, as well as communicate issues and evidence among researchers, policy-makers and public, letting us learn from, and take action based on, our mistakes. However, the manner by which adverse events are classified and reported differs from agency to agency and from treatment type to treatment type. Therefore, in order to achieve large scale integration of reports of adverse events, a careful approach to their representation must be agreed upon.

We first document the widespread agreement on what adverse events are, using examples from the Brighton Collaboration guidelines for reporting adverse events following immunization, and addresses current limitations in reporting systems that limit

their effective wider scale use. An early implementation of this approach is our Adverse Events Reporting Ontology (AERO).

2 Background

What is an adverse event?

The Guidance for Clinical Safety Data Management: Definitions and Standards for Expedited Reporting [1], defines an adverse event as “Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment.” The guide then adds “An adverse event (AE) can therefore be any unfavourable and unintended sign (including an abnormal laboratory finding, for example), symptom, or disease temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product.” The Report of Adverse Event Following Immunization (AEFI) user guide [2] from the Public Health Agency of Canada (PHAC) adheres to this

definition and adapts it for AEFI reporting: “An AEFI is any untoward medical occurrence in a vaccine which follows immunization and which does not necessarily have a causal relationship with the administration of the vaccine”.

Not detailed in this statement is the additional fact that reporting guidelines often provide protocols for determining and reporting the likelihood that specific pathological processes have occurred, and that such protocols and reporting conventions differ from jurisdiction to jurisdiction, from investigation to investigation, and by symptom and severity.

Therefore adverse events as recorded in reports, contrary to what might otherwise be presupposed, are not necessarily processes, are not necessarily of the type reports say they are, are not necessarily causally related to the intervention which led to them being reported, and the terms used to describe them are not necessarily univocal.

This matches the usage made for example within the Vaccine Adverse Event Reporting System (VAERS) [3], which asserts that “VAERS collects data on any adverse event following vaccination, be it coincidental or truly caused by a vaccine”.

In particular we distinguish adverse event from adverse side effect, where the latter is of a type determined to be causally related to the intervention.

It is a goal of our work to nonetheless provide a coherent account and workable system for managing these reports in such a way as to maximize their utility.

When are adverse events reported?

Current guidelines [2] specify that events should be reported on the basis of their temporal association with the medical intervention. For example, in the case of AEFIs depending on (i) the type of immunizing agent (30 days after live vaccine or 7 days after killed or subunit vaccine) or (ii) biological mechanism (up to 8 weeks for immune-mediated events). Even though in some cases, and based on their personal experience, clinicians may think that some adverse events are most probably caused by the intervention, and even take action based to guard the patient’s health based on this assessment, they nonetheless must report any event occurring in the respective

corresponding time frame. In that way, records accumulated from many clinicians may be reviewed by safety committees, where evidence towards causality establishment will be reviewed and policy recommendations, based on unbiased evidence, can be made.

Issues with current adverse event reporting systems

While all practitioners agree on the importance of reporting adverse events in increasing public health safety, current methods used for spontaneous adverse events reporting are not sufficient, mitigating their usefulness. For example, there is no standardization of the terminology used in the current Electronic Data Capture (EDC) used by PHAC [4]. At best, a Medical Dictionary of Regulatory Activities (MedDRA) [5, 6] code is assigned after parsing the clinician’s input, but this code is not linked to any definition. This in turn may lead to heterogeneity in the diagnoses recorded – physicians may have slightly different interpretations of what constitutes a seizure for example. Several studies highlight the potential issues in using MedDRA for adverse event reporting, ranging from inaccurate reporting as several terms are non-exact synonyms, to lack of semantic grouping features impairing processing in pharmacovigilance [7–10]. Additionally, in many systems, only the adverse event code as determined by the system (e.g. resulting from parsing the textual input) is saved, and information about signs and symptoms used in the determination of that code are lost. This limits the ability of analysts to review the set of symptoms observed in order to establish a consistent diagnosis. The resultant lack of consistency limits the ability to query and assess important safety issues the resulting datasets might otherwise support.

A general review of systems used in other countries provides similar results. The Adverse Event Reporting System (AERS) [11] and VAERS systems used in the US rely on MedDRA to encode adverse events. They follow the international safety reporting guidance [12] which specifies: “Only the MedDRA Lowest Level Term (LLT) most closely corresponding to the reaction/event as reported by the primary source should be provided”. In Europe, the Vaccine European New Integrated

Collaboration Effort (VENICE) [13] group reports [14] that only 71% (17/24) of the countries states have adopted a classification of AEFIs, and that those chosen classifications are heterogeneous: 38% WHO¹ and 62% other or not specified.

Brighton collaboration artifacts

The Brighton Collaboration [15] is a global network of experts that aims to provide high quality vaccine safety information. It has done extensive work towards standardizing the assessment and reporting of adverse events following vaccination. The Brighton Collaboration published four artifacts of interest to our current work. The case definitions they provide relate symptoms and signs to assessments of whether a particular type of pathological process has occurred, assigning qualitative levels of certainty. They provide guidelines for three activities – data collection, analysis, and presentation of results, aiming to make collected data comparable, informed by the case definitions. By determining and publishing these guidelines, the collaboration creates methodological standards that enable accurate risk assessment. The case definitions neither require, nor assess a causal relation between a given adverse event and the immunization process. Rather, the case definitions are designed to define levels of diagnosis certainty based on known information about AEFIs.

The Brighton Collaboration has published a number of papers presenting these guidelines and case definitions, each aimed at reporting potential pathological processes, such as Seizure [16] and Guillain-Barré Syndrome [17]. In our prototype we have worked with the seizure case definition.

3 Implementation

Our prototype aims to address a number of issues. While complete, the textual article-like format of the Brighton case definitions makes it difficult for clinicians to confirm that they see the relevant symptoms when making the adverse event diagnosis: case definitions are buried within the scientific paper. In the textual form, the case definition is not amenable to automated diagnosis – clinicians cannot choose which symptoms are observed and then infer the proper event diagnosis. A formal and logical description of vaccine adverse events would allow software tools to process the information and present only relevant items in a checklist to the physician, making it easier to validate upon data entry. Therefore, an ontology-based system at the time of data entry will increase data accuracy and completeness. For example, when the clinicians select seizure as adverse event, they will be offered a list of symptoms that may have manifested. By selecting the ones they did observe, the system will be able to confirm their diagnosis, potentially specifying it, such as assigning a level of certainty based on the Brighton case definition. The system will also be able to call the diagnosis into question, by warning that the set of events selected does not allow for unambiguous diagnosis. In the latter case, the system will also provide a list of such events that would allow determination. Taken together, those will enable, at the time of data entry, to unambiguously refer to a specific set of symptoms, each carefully defined, and establish a diagnosis, which remains linked to its associated symptoms. The adverse event will also be formally expressed, making it amenable to further querying for example for statistical analysis “what percentage of patients presented with motor manifestations?”) at different levels of granularity (e.g., facilitating queries such as “what percentage of patients presented with tonic-clonic motor manifestations?”) Finally, by agreeing on a common defined vocabulary we can increase data interoperability, and enable cross databases queries across different centres or against public datasets, such as literature references or other AEFI datasets.

In the following sections we present a proof of concept prototype, the AERO, based on the seizure case definition from the Brighton group.

¹ The reports can be difficult to even interpret – the WHO Adverse Reaction Terminology (WHO-ART) is a non-open terminology and only a 1997 version appears to be publicly visible, hosted at <http://biportal.bioontology.org/ontologies/40404>. It lacks many terms that are essential for AE reporting, such as those related to seizure. More importantly, WHO-ART follows a 4-level structure similar to MedDRA, and therefore suffers some of the same defects.

Incorporating existing relevant resources

In implementing a distinct resource for adverse event reporting, care was taken to reuse, when possible, work done in the context of other efforts. Reusing terms from other resources allowed us to rely on knowledge of domain experts who curated them, to dedicate more work time for terms that need to be created *de novo*. When only few terms of interest were identified in external ontologies, those have been imported relying on the Minimum Information to Reference an External Ontology Term (MIREOT) guideline [18]. For example the Vaccine Ontology (VO) [19] defines the *vaccination* process as an “administering substance in vivo that involves in adding vaccine into a host (e.g., human, mouse) in vivo with the intend [sic] to invoke a protective immune response”, and we use it to define vaccine adverse events. Similarly, we use classes from the Ontology for General Medical Science (OGMS) [20]. OGMS represents pathological entities, diseases and diagnosis, and some of its classes such as *disorder* and *sign* are at the root of very important AERO hierarchies; more details on their usage is shown below. In other cases, we decided to import external ontologies as a whole: (i) the Relations Ontology (RO) [21] contains a set of common relations, (ii) the Information Artifact Ontology (IAO) [22] deals with information entities and metadata, and (iii) the Basic Formal Ontology (BFO) is used as our upper-level ontology. Those resources are commonly used by the Open Biomedical Ontologies (OBO) Foundry [23] ontologies, of which AERO aims to be a part; relying on them for our prototype will improve integrability of our resource within the Foundry framework.

Adverse event class

Consider the following cases in which the clinician wishes to report adverse events:

- sensorineural deafness reported after measles, mumps, and rubella immunization. This disturbance of the cochlea or auditory nerve results in hearing impairment, often loss of ability to hear high frequencies [24],
- infection such as in the case of leflunomide in treatment of arthritis [25],

- any of the dermatological adverse events observed in patients treated with etanercept [26],
- headaches reported following use of proton pump inhibitors such as lansoprazole [27],
- rashes, extremely common for example at the injection site.

These cases indicate that the type of an adverse event can be either of BFO's upper level classes – *occurent* or *continuant*. OGMS currently defines *sign* as “A quality of a patient, a material entity that is part of a patient, or a processual entity that a patient participates in, any one of which is observed in a physical examination and is deemed by the clinician to be of clinical significance.” and *symptom* as “A quality of a patient that is observed by the patient or a processual entity experienced by the patient, either of which is hypothesized by the patient to be a realization of a disease.”. Those classes are siblings of the *bfo:continuant* and *bfo:occurent* classes, directly asserted under *bfo:entity*². Adverse events clearly match those definitions: they can be quality of the patient (for example, pallor or cyanosis), a material entity part of the patient (e.g., rash), or a processual entity that parts of a patient participates in (e.g., seizure).

Prototype development

Following this, we logically define *aero:adverse event* as the union of *aero:adverse event process* and *aero:disorder resulting from an adverse event process* (i.e., the adverse event *continuant* described above). An *aero:adverse event process* is “a processual entity occurring in a pre determined time frame following administration of a compound or usage of a device”; this can be logically translated (using the Manchester OWL syntax [28]) as:

² Throughout this paper we will adopt the notation *prefix:label* for entities, where *prefix* is the commonly used resource abbreviation.

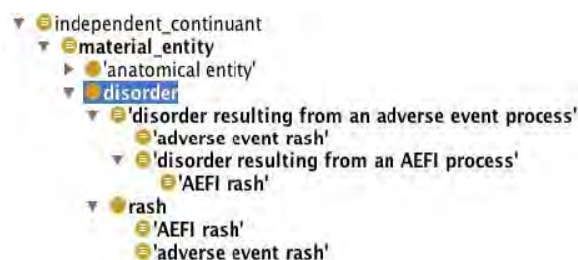


Figure 1. The disorder hierarchy as built in AERO, under the *ogms:disorder* class. The class *adverse event rash* is logically defined as the intersection of *disorder resulting from an adverse event process* and *rash*.

```
Class: 'adverse event process'
  EquivalentTo:
    processual_entity
      and (preceded_by some
        ('adding a material entity
          into a target'
            or 'administering
              substance in vivo'))
```

where the classes *adding a material entity into a target* and *administering substance in vivo* are imported from the Ontology of Biomedical Investigations (OBI) [29]. The AERO definition of *adverse event process* is meant to be inclusive, and cover cases such as those described by the Manufacturer and User Facility Device Experience (MAUDE); for example the case of a patient fitted with bioprosthetic heart valves who dies within the following 4 months³. It is also worth noting that this definition of adverse event does not imply causation between the sign observed and the compound administration/device utilization, but is rather based on temporal association.

The adverse event continuant hierarchy was built under the *ogms:disorder* class (Figure 1), which is defined as “A material entity which is clinically abnormal and part of an extended organism. Disorders are the physical basis of disease.” To avoid any language ambiguity by associating the terms event and continuant in the label of the class *adverse event continuant*, it was renamed *disorder resulting from an adverse event process*. As a general way of overcoming the potential issue between terms in use by clinicians and ontological usage in the context of the OBO Foundry, in which it may be

confusing to associate the word “event” to a hierarchical position under continuant, we chose to rely on the *OBO Foundry unique label* IAO annotation property (http://purl.obolibrary.org/obo/IAO_0000589).

Classes such as *adverse event rash* (EquivalentTo: *disorder resulting from an adverse event process* and *rash*) will therefore have an OBO Foundry unique label annotation with value “rash resulting from an adverse event process”.

Finally, as presented in the background section, adverse event definitions are based on the Brighton case definitions. We built under the IAO class *directive information entity* and defined a diagnosis guideline as a “A directive information entity that establishes a diagnosis based on a set of signs or symptoms” and its subclass *Brighton case definition* with definition “A clinical guideline in which a set of signs is described to establish a diagnosis of adverse events with a related degree of certainty, as defined by the Brighton Collaboration, <http://www.brightoncollaboration.org/>”.

The resulting inferred hierarchy, presented in figure 2, shows how defined classes are positioned under their respective parents.

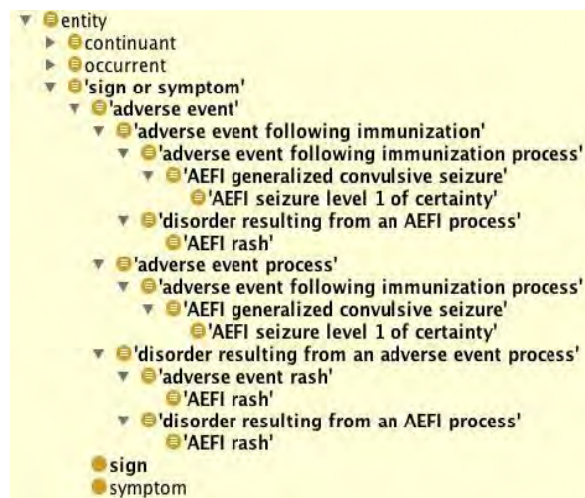


Figure 2. The resulting inferred hierarchy. Leaf terms such as *AEFI rash* are inferred under several parents based on their logical restrictions. The adverse event process class is also a subclass of *bfo:occurent*.

³ http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi_id=1942591

4 Discussion

Related efforts

As pointed out by Ceusters and al. [30], the term adverse event has been described in multiple ways. They consider adverse events as being those for which causality has been demonstrated, arguing that “the past cannot be changed: something which is not an adverse event at the time it happens cannot become one at some point thereafter.” This view is adopted by the Adverse Events Ontology (AEO)⁴, which defines adverse event as “a pathological bodily process that is induced by a medical intervention”. This definition however presents several issues that make it unusable for reporting vaccine adverse events. First, at the time of data entry, there is no certainty that the adverse event has been induced by the intervention. Instead, in clinical settings, all reactions are reported and forwarded to a committee that will later on try and establish causality. As demonstrated above, such causality will never be formerly established, rather a network of proofs is collected via epidemiologic studies and/or case reports, and supported by demonstration of biological plausibility. Second, reported adverse events may not always be processes – continuants such as rash are reported by clinicians and need to be accommodated as well. Describing the process leading to the rash is a first step as is done in the AEO, but i) this process is not always known ii) clinicians do not report nor are particularly interested in the process - they rather care about the dermatological disorder observed. Finally, that an adverse event rash and a rash non temporally associated with a medical intervention are of the same type, which should be described in a distinct, external symptom ontology - view shared by Dr. Ceusters⁵. It then becomes obvious that this rash is the universal being observed, and its nature whether we can prove or not that it is caused by the intervention – does not change. We can declare the observed rash as an adverse event on the basis of its temporal association, via a defined class as is done in AERO, describe how it follows some reporting

guidelines such as Brighton, and later on assign a category of evidence for causality, which would be an information entity attached to the original adverse event rash after review by the safety committee.

Current issues

While implementing the prototype, several ontological issues arose. Adverse event is defined as a process or continuant *preceded_by* some medical intervention or drug administration. The *preceded_by* relation is imported from the RO, and defined as “P preceded_by P’ if and only if: given any process p that instantiates P at a time t, there is some process p’ such that p’ instantiates P’ at time t’, and t’ is earlier than t”. This relation does not specify a timeframe for the events to be considered related, and an immunization process happening right after one’s birth would *de facto* precede most of the subsequent events in their life. A comment on this relation in the RO file⁶ indicates that this is an area RO developers are considering improving, and they suggest stronger relations such as *immediately_preceded_by* or an indication that the instances P and P’ share participants. Another issue related to use of relations appears when logically defining *disorder resulting from an adverse event process*. We use the relation *is specified output* to represent that each disorder results from the adverse event process. However, the range of this relation is *planned_process* – processes executed following a plan and with the intent to achieve a specific objective – which obviously adverse event processes are not. We expect BFO⁷ to provide a suitable relation for this case. Finally, there is currently no relation linking a disease and its set of signs and symptoms. This issue has been raised in OGMS⁸ but poses the question of the universality of the association between a disease and its symptoms. We currently are unable to associate all instances of the disease with the whole set of symptoms (e.g., not every instance of influenza disease is associated with an instance of fever, and certainly not vice versa). We propose here to use the case definitions (or any other source of

⁴ <http://sourceforge.net/projects/aeo/>

⁵ http://sourceforge.net/mail/archive/message.php?msg_id=27040555

⁶ <http://www.obofoundry.org/ro/ro.owl>

⁷ <http://code.google.com/p/bfo/>

⁸ <http://code.google.com/p/ogms/issues/detail?id=45>

canonical knowledge) to relate a set of signs and symptoms to a specific class, defined based on the guideline considered. For example, while not all seizures cases present with witnessed loss of consciousness (and vice versa, not all loss of consciousness are associated with a seizure episode), we can say that, *according to the Brighton case definition*, if there is a sudden witnessed loss of consciousness (in conjunction with another set of symptoms) then we can diagnose a seizure with level 1 of certainty.

Future work

Some elements of the current prototype deserve a bit more attention. For example, the levels of diagnosis certainty as defined by Brighton should probably be some type of information entity that would then be attached to the *AEFI seizure* class. Similarly, the degree of severity of adverse event, as well as their expectedness, is important information that should be modelled in the ontology. Serious adverse events⁹ are life-threatening or causing death, as well as requiring hospitalization or permanent disability. Unexpected adverse events¹⁰ are those not mentioned in drug manufacturers notices for examples. Of course, unexpected, serious, adverse event are of special concern to the safety committee. We also expect to be able to outsource our definitions of “normal” symptoms and signs. For example, the *rash* class in figure 1 should be imported from a common symptom ontology. This would however require consensus definition for those elements, which may prove difficult. We have had extensive discussion with the developers of the AEO that aided the preparation of this paper, and both groups hope we will be able to reconcile the two resources in the future.

Finally, we would like to proceed with testing of the prototype in a real use context. In the PHAC/CIHR Influenza Research Network (PCIRN) network, clinicians rely on reporting forms developed by Dacima [4]. We are in

contact with those developers to implement an extension of these forms, allowing us to maintain the interface users are already familiar with. At data entry time, they will be presented with a succession of choices augmented by their precise description and checks ensuring signs described match the adverse event reported and vice-versa. This will lead in an increased accuracy and quality of reported adverse events, and will ultimately improve patient safety.

Availability

The AERO project, including ontology and documentation which is available at:

<http://purl.obolibrary.org/obo/aero>.

AERO is also listed on the OBO library at:

<http://obofoundry.org/cgi-bin/detail.cgi?id=AERO>

and under BioPortal at:

<http://biportal.bioontology.org/visualize/45521>.

Participation in the AERO is welcome and contributions in any form are encouraged.

Acknowledgments

The authors' work was partially supported by funding from the Public Health Agency of Canada / Canadian Institutes of Health Research Influenza Research Network (PCIRN), and the Michael Smith Foundation for Health Research.

References

1. Canada Minister of Health. Clinical Safety Data Management Definitions and Standards for Expedited Reporting, <http://www.hc-sc.gc.ca/dhp-mps/prodpharma/applic-demande/guide-ld/ich/efficac/e2a-eng.php#a2A1>, June 2011.
2. Public Health Agency of Canada. User Guide: Report of Adverse Events Following Immunization (AEFI), <http://www.phac-aspc.gc.ca/im/pdf/AEFI-ug-gu-eng.pdf>, June 2011.
3. US Food and Drug Administration. Vaccine Adverse Event Reporting System Data, <http://vaers.hhs.gov/data/index>, June 2011.
4. Daciforms, http://www.dacimasoftware.com/index.php?option=com_content&view=article&id=99, June 2011.
5. MedDRA Maintenance and Support Services Organization. Medical Dictionary of Regulatory Activities, <http://www.meddrasso.com/>, June 2011.

⁹ <http://www.hc-sc.gc.ca/dhp-mps/prodpharma/applic-demande/guide-ld/ich/efficac/e2a-eng.php#a2B>

¹⁰ <http://www.hc-sc.gc.ca/dhp-mps/prodpharma/applic-demande/guide-ld/ich/efficac/e2a-eng.php#a2C>

6. Introductory Guide MedDRA Version 14.0, http://www.meddrasso.com/files_acrobat/intguide_14_0_English_update.pdf, June 2011.
7. Merrill G. The MedDRA paradox. *AMIA Annual Fall Symposium 2008*, pages 470–474, 2008.
8. Krischer J Richesson R, Fung K. Heterogeneous but standard coding systems for adverse events: Issues in achieving interoperability between apples and oranges. *Contemp Clin Trials*, 29, 2008.
9. Mozzicato P. Standardised MedDRA queries: their role in signal detection. *Drug Safety*, 30, 2007.
10. Almenoff J, Tonning J, Gould A, and Szarfman A et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Safety*, 28:981–1007, 2005.
11. US Food and Drug Administration. Adverse Event Reporting System Data, <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm>, June 2011.
12. ICH Expert Working Group E2B(R). E2B(R) Clinical Safety Data Management: Data elements for transmission of individual case safety report, <http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm129399.pdf>, June 2011.
13. Vaccine European New Integrated Collaboration Effort, <http://venice.cineca.org/>, June 2011.
14. VENICE project. Final Report on the Survey on AEFI Monitoring Systems in Member States, http://venice.cineca.org/WP5_final_report.pdf, June 2011.
15. The Brighton Collaboration, <http://www.brightoncollaboration.org>, June 2011.
16. Bonhoeffer J et al.; Brighton Collaboration Seizure Working Group. Generalized convulsive seizure as an adverse event following immunization: case definition and guidelines for data collection, analysis, and presentation. *Vaccine*, 22(5-6):557–62, 2004.
17. Sejvar JJ et al. Brighton Collaboration GBS Working Group. Guillain-Barré syndrome and Fisher syndrome: case definitions and guidelines for collection, analysis, and presentation of immunization safety data. *Vaccine*, 29(3):599–612, 2004.
18. M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg. Mireot: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33, 2011.
19. The Vaccine Ontology, <http://www.violinet.org/vaccineontology/>, June 2011.
20. Ontology for General Medical Science (OGMS), <http://code.google.com/p/ogms/>, June 2011.
21. B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005.
22. The Information Artifact Ontology (IAO), <http://code.google.com/p/information-artifact-ontology/>.
23. Ashburner M. Smith B., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L. J., Eilbeck K., Ireland A., Mungall C. J., Leontis N. OBI Consortium, Rocca-Serra P., Ruttenberg A., Sansone S. A., Scheuermann R. H., Shah N., Whetzel P. L., and Lewis S. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.
24. B J Stewart and P U Prabhu. Reports of sensorineural deafness after measles, mumps, and rubella immunisation. *Archives of Disease in Childhood*, 69(1):153–154, 1993.
25. Madhok R Alcorn N, Saunders S. Benefit-risk assessment of leflunomide: an appraisal of leflunomide in rheumatoid arthritis 10 years after licensing. *Drug Safety*, 32(12):1123–34, 2009.
26. Lidian L. A. Lecluse, Emilia A. Dowlatsahi, C. E. Jacqueline M. Limpens, Menno A. de Rie, Jan D. Bos, and Phyllis I. Spuls. Etanercept: An overview of dermatologic adverse events. *Arch Dermatol*, 147(1):79–94, 2011.
27. Angela A. M. C. Claessens, Eibert R. Heerdink, Jacques T. H. M. van Eijk, Cornelis B. H. W. Lamers, and Hubert G. M. Leufkens. Determinants of Headache in Lansoprazole Users in The Netherlands: Results from a Nested Case-Control Study. *Drug Safety*, 25(4), 2002.
28. M. Horridge, N. Drummond, J. Goodwin, A. Rector, R. Stevens, and H.H. Wang. The Manchester OWL Syntax. In Bernardo Cuenca Grau, Pascal Hitzler, Conor Shankey, and Evan Wallace, editors, *Proceedings of OWL Experiences and Directions Workshop (OWLED06)*, 2006.
29. OBI Ontology, <http://purl.obolibrary.org/obo/obi>, June 2011.
30. Ceusters W, Capolupo M, de Moor G, Devlies J, and Smith B. An evolutionary approach to realism-based adverse event representations. *Methods Inf Med*, 50(1):62–73, 2011.

Adverse Events from Clinical Studies in Pharmaceutical Research and Development

Pia Emilsson, Kerstin Forsberg

AstraZeneca R&D, Mölndal, Sweden

Abstract. A statement of interest to participate in the workshop Representing Adverse Events, arranged together with the International Conference on Biomedical Ontology (ICBO), July 2011. Introducing the business needs in handling safety issues and regular ongoing pharmacovigilance in pharmaceutical research and development. An outline of the proposed solution and two examples of different adverse event cases as a background to the authors wishes to understand a more ontological approach.

1 Business Need

R&D and the Patient Safety department have a strategic focus to develop predictive ways to handle safety issues and regular ongoing pharmacovigilance. That is, the detection, assessment, understanding and prevention of adverse effects, particularly long term and short term side effects of medicines. Regulatory authorities require well designed and proactive risk management plans to be in place from launch throughout the whole lifecycle of a product. The new IND (Investigational New Drug) regulation from FDA, for routine review of incidence rates of all serious and non-serious adverse events in all clinical programs.

Many clinical study programs run several studies worldwide in parallel, which could result in a high worldwide exposure; hence it is extremely important to have continuous access to ongoing clinical study data. To successfully handle ongoing pharmacovigilance a prerequisite is access to continuous relevant pooled clinical study data in a format that make it possible to review, search and answer questions in a very short time frame. This requires consistent coding, pre-prepared pools and derived variables. Furthermore, the result of the searches in the pooled clinical study data should be put in context of other results both inside and outside the pharm company. In addition it is required to keep track of each single data point through the data collection and

refinement chain as this contributes to the final result.

The overall business problem as stated for the ongoing AstraZeneca project called Quest: No global automatic way to access, structure and analyze safety related clinical study data (including ongoing study data) at drug product/project level.

Below an outline of the proposed solution for representation of adverse events influences by existing standards focused on data exchange and coding. This overview and two examples of cases provide some background to our interest in better understanding a more ontological approach to represent adverse events in the context of pharmaceutical research and development. We also explore semantic web standards and linked data principles to improve the research utility of data in clinical studies.

2 Solution Overview

A safety analysis environment has been proposed by the Quest project. It includes: Clinical study data (legacy, completed, ongoing), Structure (optimized for Patient Safety queries), Analyze – (descriptive statistics and graphical output), Processes (how to work using these new options).

A relational database design, building on the experiences of two large Adverse Events databases in production for approximate 20 years have been proposed for the central Quest database.

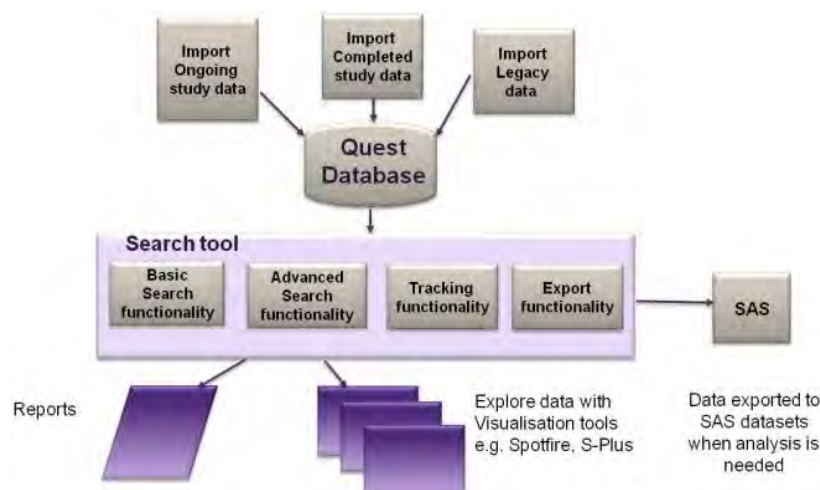


Figure 1. Proposed Quest solution

A key concept in the design of the Quest database is periods e.g. run-in, on treatment, wash-out, follow-up and the sequence of these periods. The periods will also reflect the half-life of the drug studied. All findings, events and interventions need to be linked to one or several periods.

This is also an example of how existing data exchange standards have influenced the mindset in pharmaceutical research and development on how adverse events should be represented. In this case it is the general classification of “observations” i.e. data exchange records, in terms of findings, events and interventions according to CDISC’s (Clinical Data Interchange Standards Consortium) standard called SDTM (Study Data Tabulation Model). Another example is SDTM standard for data structures and elements, e.g. the extended AE domain for safety analysis. Together with CDISC’s controlled terminology with lists of codes (text strings to be used as submission values) for safety analysis such as the Severity/Intensity Scale for Adverse Events, with text strings: “MILD”, “MODERATE”, “SEVERE”.

The dictionary for coding of explicit AE records in clinical trials sponsored by pharmaceutical companies is MedDRA. Adverse events can also be the implied consequence of a combination of measurements. Below two examples issues when it comes to versioning of MedDRA terms and an example of a lab measurement based adverse event case.

Issue #1

A main issue is the use of different MedDRA versions with potentially new so called preferred terms (PT:s) and new hierarchies. However no low level terms (LLT) are deleted/reused.

Original Coded MedDRA LLT	Original Coded MedDRA PT	Current MedDRA PT
Heaviness of head	Headache	Head discomfort
Headache vasomotor	Headache	Cluster headache
Headache Unilateral	Headache	Hemicephalgia

Figure 2. Example of MedDRA terms and versions

Issue #2

Adverse events can also be the consequence of a combination of measurements. For example FDA’s document “Guidance for Industry Drug-Induced Liver Injury: Premarketing Clinical Evaluation” from July 2009 describes the Hy’s Law (PHL).

- A Potential Hy’s Law (PHL) case is defined as any situation where a study subject
 - Has an increase in both alanine or aspartate aminotransferase (ALT or AST) $\geq 3 \times \text{ULN}$ and total bilirubin (TBL) $\geq 2 \times \text{ULN}$
 - Irrespective of alkaline phosphatase (ALP), at any point during the study
 - The elevations do not have to occur at the same time or within a specified time frame

Classifying Adverse Events from Clinical Trials

Bernard LaSalle, Richard Bradshaw

University of Utah, Biomedical Informatics, Salt Lake City, UT, USA

Abstract. The use of adverse event data from investigator-initiated clinical trials, outside of the study event itself, has not been practical because of the absence of uniform data collection. The release of CTCAE 3.0 provided a controlled vocabulary for several components of adverse event data, but was difficult to use in a research setting. We have created a web-based application that combines data from clinical trials and the CTCAE classification information into a queryable database.

1 Introduction

The pharmaceutical and device research community has been actively involved in collecting, classifying and reporting adverse events (AEs) since the advent of efforts to improve human subject protection [1,2]. Within this domain there has been acceptance of terminology standards for AEs. However, investigator initiated clinical research makes up the majority of clinical research conducted within the United States [3] and within this domain use of standard terminology and classification is not a common practice.

Initial and follow-up reporting of AEs during clinical trials is primarily driven by the regulatory process, which is determined by the requirements of the sponsor, the Food and Drug Administration (FDA) [4] and the Office of Human Research Protection (OHRP) [5]. The data collected are used to determine if the AE is serious, if the AE affects the enrollment status of the study participant, and whether the AE is related to the study drug, device or procedure. Once this process is complete, formal classification of the adverse event is required for reports to data and safety monitoring committees and as part of the statistical analysis plan.

We have created a straightforward, browser-based application to provide domain experts enough information to classify AEs using the National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE) [6], which is an initial step in the process of designing an ontology for clinical research AEs.

2 Background

Investigator-initiated clinical research in the United States that involves human participants routinely includes processes for identifying AEs and collecting data about the event. Frequently the person(s) collecting these data are clinical coordinators or study nurses and not the principal investigator. When AEs are considered to be serious, a mandatory reporting process is usually started. This process notifies various regulatory entities (IRB, DSMB, sponsor, FDA) [7-9] with a minimum data set of information that fulfills the short-term requirements for reporting, but does not provide enough metadata about the event to be useful outside the context of the initial report. AEs that are determined to be non-serious may not receive any further evaluation other than their frequency.

Classifying AEs in a standardized way (system organ class, term, definition, grade and severity) requires someone with domain knowledge of the clinical presentation, the study protocol and the classification standard. Correlating and evaluating all of these data is time consuming, requiring simultaneous review of multiple data sources.

3 Objective

Our objective was to create a web-based application that could import AE data from a clinical or device study and display that data, related study data and a search window of CTCAE data within a single browser page. This display allows a domain expert to quickly evaluate the AE data, determine the correct

CTCAE classification and then, with a few mouse clicks, link these data and store them in a new AE classification table which is associated with the study meta data.

4 Methods

Since the application is database driven, we needed a data model that could import research data with a minimum of extraction-transpose-load (ETL) effort. Only meta data about the study and a subset of the AE data are imported into two tables (Figure 1). This is currently a manual process using a formatted text export file or a direct SQL query of the study database.

The published NCI CTCAE data are loaded into multiple tables by a similar process; however, since this is a known structure, the process is straightforward and required only when CTCAE data are updated. The CTCAE data are loaded into three tables in a relational database, PostgreSQL [10], to enable faster searching by the person evaluating the AEs (Figure 2).

The application interface allows the user to select a study that filters the display of AE data. As each AE is displayed, the user can select a System Organ Class (SOC), filtering the number of AE terms displayed, or an AE term directly. Once the Preferred Term has been selected, the corresponding Grades are displayed. Once the Grade has been chosen, all of the data from both data sources are ready to be inserted into the Evaluated AE table (Figure 3).

5 Results

The application allows users to view AEs in a browser interface within a single device screen. Users have the ability to view only AEs that require classification, filtered by study. A search engine allows users to reduce the focus of data being displayed by selecting either the SOC or the term relevant to the AE being classified. Once the user has determined the correct classification, she then selects the correct grade. The associated classification data

and AE data from the study are automatically populated into a data table. Previously classified AEs can be reviewed and edited within the same interface. Meta data about the study are assigned a unique identification number that is associated with each classified AE. This allows comparison of AEs within a single trial or across trials when appropriate.

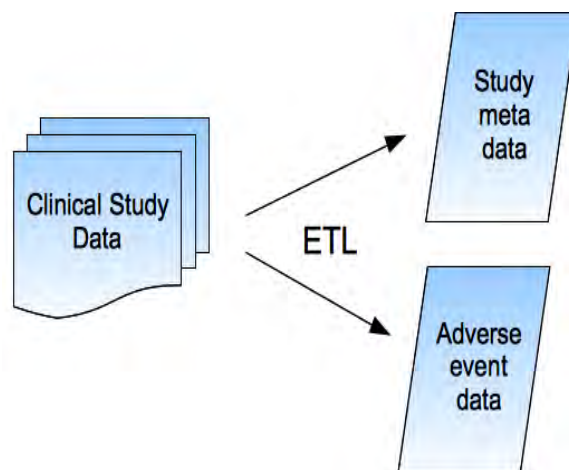


Figure 1. The loading of clinical study data to be classified into the application database tables. Only data relevant to classifying the AE are loaded.

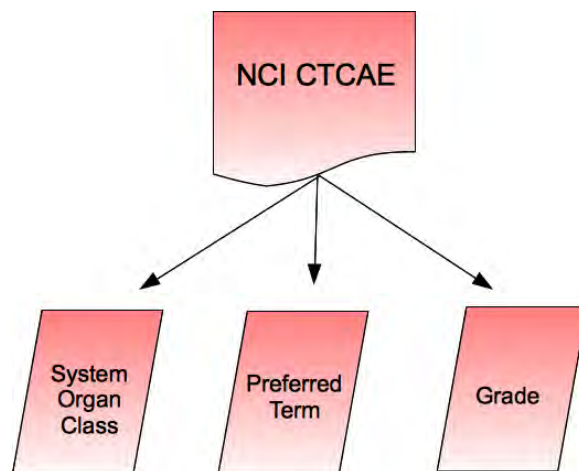


Figure 2. The reorganization of CTCAE data into relational tables. Each System Organ Class contains the AE preferred terms within the class and each term has several possible Grades.

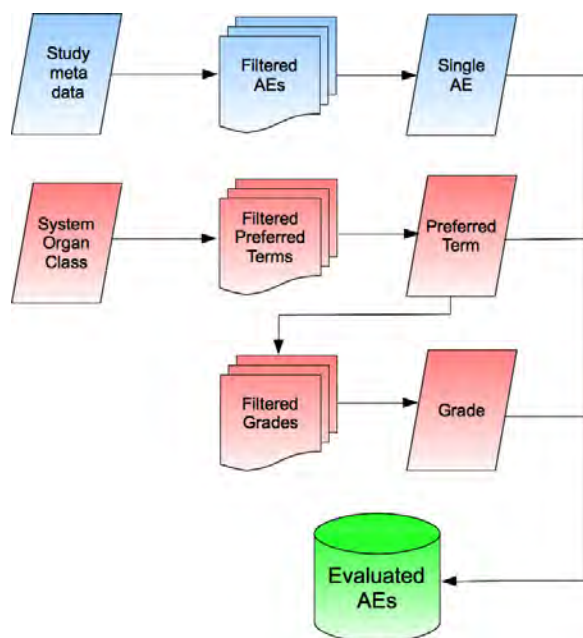


Figure 3. The process flow 1) Selecting a study and a single AE from the subset of AEs filtered by the study ID; 2) Selecting the SOC and a single preferred term filtered by the SOC; 3) Selecting a grade from the grades associated with the preferred term. The data from these three steps are merged and stored.

6 Conclusion

Individual research sites and data centers will use a variety of data models and data sources when collecting clinical research data. We have created an application that simplifies importing AE data through the use of mapping tables during the ETL process. By using an existing classification standard (CTCAE) and a standard results table, we can standardize the AEs from multiple trials once they are classified. The resulting data can now form the basis for discussions about whether CTCAE is the most appropriate classification system, whether the meta data collected are sufficient to compare AEs across trials, and how this could contribute to the creation of an ontology for AEs.

The classifying and reuse of adverse event data can be a significant contribution to patient safety, and is critical to orphan disease clinical

research, particularly within pediatric populations. Standardized vocabularies and concepts of AEs may allow investigators to reuse safety data from previous clinical research, which would significantly reduce the cost and length of clinical research trials.

References

1. The National Research Act, United States Congress, 1974.
2. The Belmont Report, Ethical Principles and Guidelines for the protection of human subjects of research, Department of Health, Education and Welfare, 1979
3. The Investigator-sponsored IND in clinical trials, Haakenson C, Fye CL, Sather MR, Toussaint DJ, Control Clin Trials. 1987 Jun;8(2):101-9.
4. The Food and Drug Administration, Department of Health and Human Services, United States of America, 1930.
5. Office of Human Research Protection, Department of Health and Human Services, United States of America, 1966.
6. Common Terminology Criteria for Adverse Events, National Cancer Institute, National Institutes of Health, United States of America, August 9, 2006.
7. Institutional Review Board, National Research Act of 1974, USA, are governed by Title 45 CFR (Code of Federal Regulations) Part 46, National Institutes of Health, Department of Health and Human Services, United States of America.
8. Data and Safety Monitoring Board (DSMB), NIH Policy For Data and Safety Monitoring, National Institutes of Health, United States of America, June 10, 1998.
9. Sponsor, The entity responsible for registering the study is the "responsible party.", as defined in 21 CFR (Code of Federal Regulations), Part 21 CFR 50.3, National Institutes of Health, Department of Health and Human Services, United States of America.
10. *PostgreSQL 9.0.0 Documentation*. PostgreSQL Global Development Group. Retrieved 2010-09-20.

An Ontological Representation of Adverse Drug Events

Guoqian Jiang¹, Jon D. Duke², Jyotishman Pathak¹, Christopher G. Chute¹

¹Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, USA

²Regenstrief Institute, Indianapolis, IN, USA

Abstract. A standardized, controlled vocabulary allows adverse drug events (ADE) information to be described in an unambiguous way in knowledge bases, which is critical for clinical decision support systems for patient medication safety. In this paper, we describe our preliminary effort on development of an ontological representation pattern for the ADE domain. We discuss clinical implications of the effort and potential challenges on its integration with existing data standards.

1 Introduction

Adverse drug events (ADE) are a well-recognized cause of patient morbidity and increased health care costs in the United States. Multiple studies have demonstrated that a clinical decision support (CDS) system based on a standardized ADE knowledge base can be useful to help physicians reduce the risk of their patients' medications [1,2]. In a previous study, for instance, we proposed a comprehensive framework for building a standardized ADE knowledge base known as *ADEpedia* (<http://adepedia.org>) through combining ontology-based approaches with Semantic Web technology [3]. However, there is no standardized, controlled vocabulary available that allows the ADE information to be described in an unambiguous way in such a knowledge base.

An ontological representation of the ADE domain would provide computable semantics for an ADE knowledge base, and facilitate semantic integration of ADE related data standards. In the present work, we describe our preliminary effort on development of an ontological representation pattern for the ADE domain. We discuss clinical implications of the effort and potential challenges with respect to its integration with existing data standards.

2 Related Work

Stetson, *et al.* (2001) developed an ontology representing the intersection of medical errors,

information needs and the communication space [4]. The main use of that ontology was to help guide the rational deployment of informatics interventions. Herman, *et al.* (2005) created a vaccine adverse event ontology for public health [5]. Mokkarala, *et al.* (2008) described their efforts in developing a comprehensive medical error ontology to serve as a standard representation for medical error concepts from various existing published taxonomies [6]. Ceusters, *et al.* (2011) described an evolutionary approach to realism-based adverse event representations [7]. The ontology is designed under the OBO foundry principles and merged with the Basic Formal Ontology (BFO) in upper level. Although these ontology development efforts are relevant to the ADE domain and would be useful starting points, the semantics of the ADE domain remain poorly specified. For example, in the Adverse Event Ontology (AEO) [8], there are only two ADE related concepts defined: the concept "drug adverse event" under its parent "adverse event" and the concept "drug administration" under its parent "medical intervention". For another example, Bousquet, *et al.* (2008) [9] proposed an ontological model for adverse drug reactions, in which the main categories "Investigation", "Accident" and "Disorder" are related to "Drug" with the "Is_related_to" link.

3 An Ontological Representation Pattern of the ADE

We proposed an ontological representation pattern for the ADE domain. We also

performed a case study for representing real ADE data using the domain pattern.

Fig. 1 shows the proposed pattern for the ADE domain. In the pattern, we defined four major types: Adverse Drug Effect Class, Drug Class, Adverse Drug Effect and Medication. We also defined the relationships between the four types. The relationships are “may_induce”, “may_be_induced_by”, “has_member” and “is_member_of”. Table 1 shows the relationship definition using the vocabulary of RDF Schema (RDFS) [10].

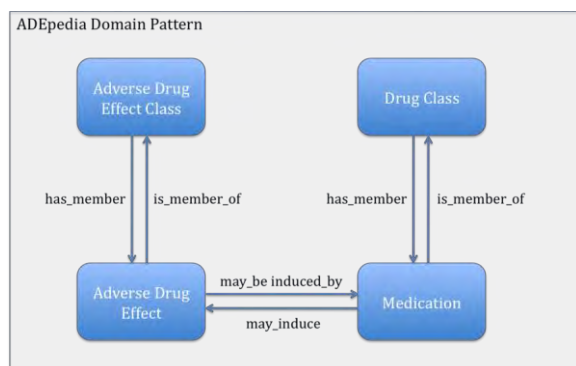


Figure 1. Proposed pattern for the ADE domain

With the pattern specified, we will be able to represent real ADE data. Table 2 shows the example instances for each of the four types defined in the pattern. An ICD-10-CM [11] code “*Adverse effect of anticoagulants*” is used as an instance of the type ADE Category; a NDF-RT (National Drug File-Reference Terminology) [12] code “*Anticoagulants*” as an instance of the type Drug Class; a SNOMED CT [13] code “*Blood in urine*” as an instance of the type ADE; and a RxNORM [12] code “*Warfarin sodium*” as an instance of the type Medication.

4 Discussion

This study was motivated by our ongoing ADEpedia project aiming to develop a standardized ADE knowledge base. We have extracted both medication and ADE data using FDA Structured Product Labels (SPL) [14]. In the knowledge base, the medication data are represented by RxNORM codes and the ADE data are represented by SNOMED CT and MedDRA [15] codes. We use “may_induce” and “may_be_induced_by” to

represent the drug-ADE relationship. We chose the predicates because they cover uncertainty between a drug and an ADE. For example, “Severe nausea occurs in 20% of patients for a given medication”. We are also exploring the approach to represent this kind of frequency and severity knowledge of the ADEs.

From the clinical perspective, clinical decision support (CDS) rules related to drugs and ADEs are generally expressed using a therapeutic or pharmacologic class (e.g. ACE Inhibitors) or a class of ADEs (e.g. adverse effects on cardiovascular system), rather than an individual drug or an ADE. Therefore, we plan to aggregate the medications to drug classes and the ADEs to ADE classes for our ADEpedia knowledge base.

In the proposed pattern, we use “has_member” or “is_member_of” to explicitly represent the drug-class membership relation. Note that the relation between individual drugs and drug classes or between individual ADEs and ADE classes is generally represented through a taxonomic relation (isa) in biomedical terminologies [16]. Usually, the isa relation is translated to the “rdfs:subClassOf” in the vocabulary of RDF Schema.

```

<adepedia:may_induce> a <rdf:Property>;
  <rdfs:domain> <adepedia:Medication>;
  <rdfs:range> <adepedia:AdverseDrugEffect> .
<adepedia:may_by_induced_by> a <rdf:Property>;
  <rdfs:domain> <adepedia:AdverseDrugEffect>;
  <rdfs:range> <adepedia:Medication> .
<adepedia:has_member> a <rdf:Property>;
  <rdfs:domain> <adepedia:AdverseDrugEffectClass>;
  <rdfs:range> <adepedia:AdverseDrugEffect> .
<adepedia:is_member_of> a <rdf:Property>;
  <rdfs:domain> <adepedia:AdverseDrugEffect>;
  <rdfs:range> <adepedia:AdverseDrugEffectClass> .
<adepedia:has_member> a <rdf:Property>;
  <rdfs:domain> <adepedia:DrugClass>;
  <rdfs:range> <adepedia:Medication> .
<adepedia:is_member_of> a <rdf:Property>;
  <rdfs:domain> <adepedia:Medication>;
  <rdfs:range> <adepedia:DrugClass> .

```

Table 1. Relationship definition using RDF Schema

```

_:b0 a <adepedia:AdverseDrugEffectClass>;
    <adepedia:code> "T45.515";
    <adepedia:displayName> "Adverse effect of
    anticoagulants";
    <adepedia:codeSystemName> "ICD-10-CM" .

_:b1 a <adepedia:DrugClass>;
    <adepedia:code> "C8812";
    <adepedia:displayName> "Anticoagulants";
    <adepedia:codeSystemName> "NDF-RT" .

_:b2 a <adepedia:Medication>;
    <adepedia:code> "114194";
    <adepedia:displayName> "Warfarin Sodium";
    <adepedia:codeSystemName> "RxNORM" .

_:b3 a <adepedia:AdverseDrugEffect>;
    <adepedia:code> "34436003";
    <adepedia:displayName> "Blood in urine";
    <adepedia:codeSystemName> "SNOMED CT" .

```

Table 2. Example instances of the four major types of the ADE representational pattern in RDF Turtle format.

In addition, we consider existing ADE relevant data standards can provide standardized codes as the instances (or subtypes) of the four types defined in the pattern. For instance, Bodenreider, *et al.* (2010) investigated drug classes in biomedical terminologies from the perspective of clinical decision support. For 134 target drug classes, SNOMED CT was identified as the single best source with 75% coverage [16]. Pathak, *et al.* (2010) analyzed categorical information in two publicly available drug terminologies: RxNorm and NDF-RT [12]. For the ADEs and ADE classes, we consider that SNOMED CT, MedDRA [15] and ICD [11] will be good candidate sources for further investigation.

In summary, we identified and defined a domain pattern for the ADE knowledge representation and we consider this pattern can be a starting pointing for an ontological representation of ADE domain. We believe that a community-based effort would be required to achieve a comprehensive standardized ontology for the domain, which would facilitate the semantic interoperability of the ADE knowledge bases in heterogeneous CDS systems and ultimately improve patient safety.

Acknowledgments

This study is supported in part by the Mayo Clinic Center for Clinical and Translational Research (CTSA) grant (RR 24150) and the

Pharmacogenomics Research Network (PGRN) Ontology Network Resource grant (GM 61388).

References

1. Duke JD, Friedlin J. ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. AMIA Annu Symp Proc. 2010 Nov 13;2010:177-81.
2. Duke JD, Li X, Grannis SJ. Data visualization speeds review of potential adverse drug events in patients on multiple medications. J Biomed Inform. 2010 Apr;43(2):326-31.
3. Jiang G, Solbrig HR, Chute CG. ADEpedia: a scalable and standardized knowledge base of adverse drug events. AMIA Annu Symp Proc. 2011. (in submission)
4. Stetson PD, McKnight LK, *et al.* Development of an ontology to model medical errors, information needs, and the clinical communication space. Proc AMIA Symp. 2001:672-6.
5. Herman TD, Liu F, Sagaram D, *et al.* Creating a vaccine adverse event ontology for public health. AMIA Annu Symp Proc. 2005:978.
6. Mokkarala P, Brixey J, Johnson TR, *et al.* Development of a Comprehensive Medical Error Ontology. In: Advances in Patient Safety (Vol. 1): Agency for Healthcare Research and Quality; 2008.
7. Ceusters W, Capolupo M, de Moor G, *et al.* An evolutionary approach to realism-based adverse event representations. Methods Inf Med. 2011;50(1):62-73. Epub 2010 Nov 8.
8. The AEO: <http://www.aeo-ontology.org/browser/index.php?o=AEO>.
9. Bousquet C, Trombert B, Kumar A, Rodrigues JM. Semantic categories and relations for modelling adverse drug reactions towards a categorial structure for pharmacovigilance. AMIA Annu Symp Proc. 2008 Nov 6:61-5.
10. RDF Schema: <http://www.w3.org/TR/rdf-schema>
11. <http://www.cdc.gov/nchs/icd/icd10cm.htm>.
12. Pathak J, Chute CG. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. J Am Med Inform Assoc. 2010 Jul-Aug;17(4):432-9.
13. The IHTSDO: <http://www.ihtsdo.org/snomed-ct>
14. The FDA Structured Product Labeling <http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm>.
15. The MedDRA: <http://www.meddrasso.com/>.
16. Bodenreider O, Fushman DD. Investigating drug classes in biomedical terminologies from the perspective of clinical decision support. AMIA Annu Symp Proc. 2010:56-60.

Toward Answering Time-Related Questions from Adverse Event Reports Using Ontology-Based Approaches

Cui Tao¹, Guoqian Jiang¹, Kim Clark^{2,3}, Deepak Sharma¹, Christopher G. Chute¹

¹Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic,
Rochester, MN, USA

²Department of Biomedical Engineering, University of Minnesota, Minneapolis, MN, USA

³Boston Scientific Corporation, Maple Grove, MN, USA

Abstract. Identifying the time patterns from medical device adverse event reports can help clinical researchers or medical devices manufacturers to understand cause of events and predict future events. Ontology-based approaches can provide a reliable and effective system for representing both events and temporal relations, annotating the useful information from narratives, and querying and reasoning over the data. This paper discusses our vision on such a system, as well as challenges we encountered during our preliminary studies.

1 Introduction

Potential temporal patterns may exist within reports of similar adverse events or similar medical devices. These patterns may include a similar sequence of events, durations of or between events, or a time/date during which the adverse event occurred. These temporal properties and relationships, however, are often buried within the text of the narrative, requiring an astute observer to detect patterns while reading several reports for the same failure mode. This method for assessing tens of thousands of adverse event reports is time consuming, expensive, and the potential exists for a missed pattern observation or error in interpreting event sequencing. In addition, because temporal relations may require inference if they are not explicitly expressed within the narrative, temporal reasoning is also needed in order to analyze the trends in time. An automated temporal analysis of reports for similar adverse events across similar products of multiple device manufacturers could lead to faster identification of trends, quicker identification of the origin of the adverse event, a more detailed understanding of the events leading up to the failure, and earlier prediction of a future failure based on similarities in event order and/or duration.

In this paper, we introduce our vision on an ontology-based system for answering time-related questions based on adverse event reports from the Food and Drug Administration (FDA) Manufacturer and User Facility Device Experience (MAUDE) database, which contain more than 1.64 million files as of March of 2011 [4]. Figure 1 shows the system overview. The arrows denote the system workflow. The Hexagons indicate important steps for 1) processing the data: selecting the adverse reports from the MAUDE database, 2) annotating the reports with respect to domain ontologies, and 3) retrieving answers of time-related questions by querying and reasoning over the annotated data. In the rest of the paper, we introduce our preliminary studies of such a system and discuss lessons learned for improving the system.

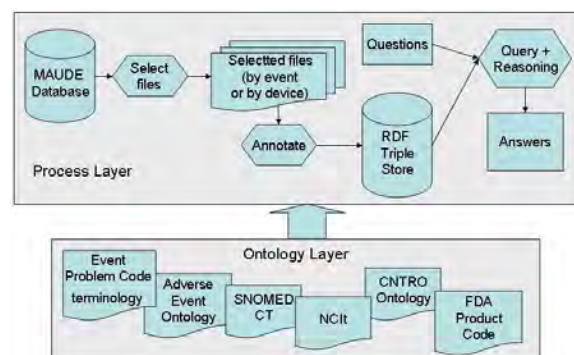


Figure 1. System Overview

2 Approaches and Challenges

File Selection for Devices

The MAUDE database provides a way to search files using the FDA product code [3]. Currently the product codes are maintained in a relational database: FDA Product Classification Database. It contains device names and their associated product codes, which identify the generic category of a device for FDA. The database, however, does not contain much hierarchical information of the devices, nor does it contain annotation properties such as synonyms, acronyms, and etc. This makes it difficult to find the right product code(s) for a specific category of devices. For example, to find the code for drug-eluting stent, one has to search by the exact phrase “drug-eluting stent”, synonyms such as “drug coated stents” or “medicated stents”, or acronym such “DES” would not work. In addition, the product codes do not contain much hierarchical classification. It is hence also difficult to find a category of devices when on a different levels of granularity are used in the classification by FDA. For example, the FDA classification does not provide a category for ventricular assist device (VAD), but rather contains codes for different kinds of VADs. This makes it difficult for computer programs to automatically identify all the codes for VAD without any domain knowledge. We believe a domain ontology for medical device classification is very necessary for the purpose of our research. The ontology can be created on the basis of the FDA Product Classification Database, with extended hierarchical information, and annotation properties.

File Selection for Events

The second challenge we encountered is to identify similar kinds of adverse events. The MAUDE database does not provide any searchable feature for adverse events. Currently one has to manually go through the narratives to identify a specific adverse event. A domain ontology would help automate this process. The Adverse Event Ontology (AEO) is a community-driven ontology developed to standardize and integrate data on biomedical adverse events and support computer-assisted reasoning [1]. The current version of AEO,

however, focuses on vaccine and drug adverse events as the initial use cases, and only provides a shallow representation for medical device adverse events. It will be an interesting and important task to extend the AEO for representing medical device adverse events. In addition, since our focus is the time aspects in adverse events, we would also like to assert the temporal information in adverse events semantically when possible (e.g. to assert that a late stent thrombosis is a thrombosis that has a duration of 30 days or longer).

Semantic Annotation

Since a lot of temporal information is embedded in clinical narratives, extracting information of interest and semantically annotating this information with respect to domain ontologies are a must. A manual annotation process could be very time-consuming and expensive [7]. The information to be annotated for our system includes: adverse events, other related medical events, possible device problems, as well as important temporal expressions and relations. We are trying to adapt NCBO annotator [5] and Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) [6] for semi-automatic annotation in our system. The NCBO annotator is a user-friendly, scalable web service that is designed for annotating narratives with relevant ontology concepts. The cTAKES pipeline processes clinical notes, identifies important clinical named entities (NEs) as well as semantic features such as negations for the NEs, and assigns attributes for the code in standard ontologies such as SNOMED CT [9] and RxNorm [8] to these NEs. Most recently, a time relation recognizer is under development for cTAKES. Both systems, however, rely on well-developed domain ontologies as the background knowledge base for annotation. Currently, there is a lack of a domain ontology that can serve the purpose as a knowledge base for medical device adverse events. The FDA has developed a set of Adverse Event Problem Codes for standardizing the classification of device and patient problems associated with medical device use [10]. This coding system can provide a good foundation for developing an ontology of the domain. A few extensions, however, need to be done to serve the annotation purpose: (1)

expanding hierarchical levels to cover granular representation of data, and (2) adding more lexical properties for each code, i.e., synonyms and acronyms.

Reasoning Framework

We are working on building an ontology-based temporal relation reasoning framework for helping clinical researchers answer time-related questions from EHR. We have built an ontology called CNTRO (Clinical Narrative Temporal Relation Ontology) [12], which was designed to model the temporal information in clinical narratives. It models clinical events, different kinds of temporal expressions (such as time instants, time intervals, repeated time periods, and durations), different levels of time granularity, temporal relations, and time uncertainties. Based on CNTRO, we have also developed a prototype framework for querying and inferring temporal information [11]. We evaluated the feasibility of using CNTRO with existing Semantic-Web technologies and discussed possible limitations and extensions that we found necessary or desirable to achieve the purposes of querying time-oriented data from real-world clinical narratives. Most recently, we have applied our system to the late stent thrombosis adverse events use case (with manually annotated data). Our preliminary study received very promising results (~89% accuracy) on answering important time-related questions [2].

3 Concluding Remarks

In this paper, we discuss our vision of an ontology-based system for representing both events and temporal relations, annotating the useful information from narratives, and querying and reasoning over the data. Our preliminary study indicates that our system is feasible for retrieving from and reasoning over annotated data from medical device adverse event reports. Future work includes extending or creating domain ontologies for medical

devices, adverse events associated with medical devices, as well as problems associated with medical devices.

References

1. Adverse Event Ontology, <http://sourceforge.net/projects/aeo/>.
2. K.K. Clark, C. Tao, D.K. Sharma, and C.G. Chute. Application of a temporal reasoning framework tool in analysis of medical device adverse events. In AMIA Annual Symposium, 2011 (submitted).
3. FDA product code, <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpdc/classification.cfm>.
4. MAUDE – Manufacturer and User Facility Device Experience Database – FDA, <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm>.
5. NCBO Annotator, <http://www.bioontology.org/annotator-service>.
6. cTAKES, <http://www.ohnlp.org>.
7. Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. Building a semantically annotated corpus of clinical texts. *J. of Biomedical Informatics*, 42:950–966, October 2009.
8. RxNorm, <http://www.nlm.nih.gov/research/umls/rxnorm/>.
9. Systematized nomenclature of medicine-clinical terms (SNOMED CT), <http://www.snomed.org>.
10. L. Reed TL T and D. Kaufman-Rivi. FDA adverse event problem codes: standardizing the classification of device and patient problems associated with medical device use. *Biomedical Instrumentation and Technology*, 44(3):248–256, 2010.
11. C. Tao, H.R. Sobrig, D.K. Sharma, W.-Q. Wei, G.K. Savova, and C.G. Chute. Time-oriented question answering from clinical narratives using Semantic-Web techniques. In ISWC, pages 241–256, 2010.
12. C. Tao, W.-Q. Wei, G.K. Savova, and C.G. Chute. A semantic web ontology for temporal relation inferencing in clinical narratives. In AMIA Annual Symposium, pages 787–791, 2010.

Supporting Medical Device Adverse Event Analysis in an Interoperable Clinical Environment: Design of a Data Logging and Playback System

David Arney¹, Sandy Weininger², Susan F. Whitehead¹, Julian M. Goldman^{1,3}

¹MD PnP Program CIMIT, Cambridge, MA; ²U.S. Food and Drug Administration CDRH/OSEL/DESE;

³Departments of Anesthesia and Biomedical Engineering, Massachusetts General Hospital, Boston, MA

1 Introduction

It is often difficult or impossible for clinicians or regulators to find root causes of failures when adverse events happen. This problem will only get worse as healthcare organizations integrate more and more devices into their information systems in order to accomplish meaningful use of electronic medical records and to meet objectives for improved patient safety [3]. Moreover, the risk of liability is one of the factors in the reluctance of medical device manufacturers to make their devices interoperable. [4] This is becoming increasingly relevant to health care providers due to the recent FDA ruling on Medical Device Data Systems (MDDS) [5], which will require applicable hospitals to register as device manufacturers. Risk management standards including ISO/IEC 80001 [6] require users to mitigate many of the risks associated with interconnecting medical devices. A data logging and playback system can address many of these needs. Data loggers for interoperable systems should capture commands, device connections and disconnections, physiologic and technical alarms, physiologic data from patients, and other information about the status of devices. In this paper, we explore the issues involved in designing such a logging system and present some preliminary solutions.

Healthcare delivery organizations need a data logger for network-integrated devices that can capture the data needed for effective adverse events analysis. The purpose of the data logger is to record low-level device data (e.g. button presses and physiological data values) from individual medical devices, along with location information and data about the status of the medical device network, in an open, standardized, and time-synchronized manner. It is impossible to trace back to the origins of

interactions between devices that can cause serious hazards to patients without a coordinated, time-synchronized log of all of the data sent by all of the devices in the system. This complete data record offers a more complete event picture than the highly filtered and processed data that goes into the EMR.

Challenges must be overcome at each step of adverse event analysis. Simply locating devices is frequently difficult, and unless devices are immediately sequestered following an incident their internal data log may be overwritten or deleted. Much data is entered manually, raising the problem of retrospective documentation where the clinician enters a value from memory or enters what they reconstruct it might have been. Reported times in the records may come from the clock on the wall, a device, or the clinicians wristwatch; current recording systems are not time synchronized. Thus, even something as simple as the start time of surgery or the time an infusion was started may be different in the nursing record versus the device internal log versus the anesthesia record.

Analysis of medical adverse events is commonly a manual process, whether a patient is affected or the event is an unexpected device interaction that does not directly impact patient care. The analysis team must locate and sequester the devices involved, get data out of the devices in a format they can use, analyze the data to figure out what happened, and then produce reports detailing their findings. Interviewing of clinicians is an integral part of the process. Adverse event reports are usually captured individually, on a case-by-case basis. Hazard analyses of medical devices and systems are based on collections of these reports, and the safety of devices is predicated on these hazard analyses. Adverse events may occur even if all of the devices act in accordance with their

specifications; if this happens repeatedly it may be an indication that something needs to change – possibly the device specifications, other processes, or the device’s use environment. After an event is analyzed and the root cause is found, it is usually impossible to know how widespread the problem is. Because detailed logs are not kept, is not possible to look back at similar situations in the past to see if a similar chain of events occurred that could indicate other undetected adverse events or near misses. Adverse event analysis thus must operate on a series of disconnected, anecdotal, individual cases rather than being able to apply epidemiological principles to consider device failures across populations.

A recorder that logs data from all of the medical devices attached to a patient has the potential to facilitate radical improvements in patient safety, with the added benefit of simplifying troubleshooting of network-related problems. Event logs and adverse event analysis entail a cross-cutting effort across clinical engineering, IT, compliance, biomedical engineering, quality assurance, and clinical care in the OR, ICU, and other settings. We believe that better collection and more accurate documentation of adverse events will lead to safer medical devices and systems in the future.

2 Design

Our design is based on the ICE (Integrated Clinical Environment) architecture from the ASTM F2761-09 standard [2]. Medical devices associated with a single, high-acuity patient are all connected through an ICE Network Controller that contains a data logger. Figure 1 shows the general architecture of an ICE system. As medical device interface capabilities improve, more device data will be available to the data logger. F2761-09 requires logging of “user interaction with devices” – e.g., button presses – that will help add context to events to facilitate analyses of usability problems.

Types of Logging

The event recorder will be useful for analyzing adverse events and near misses with patients as well as debugging interactions between multiple medical devices (such as bedside monitors and remote alarm systems) or between medical devices and other IT systems (e.g. the EHR). We

anticipate that this data will also be extremely useful for developing advanced clinical algorithms and analyzing patient outcomes. Log data for debugging network interactions will typically be much more detailed than that used for clinical event analysis. For instance, when a pulse oximeter transmits SpO₂ on the network, a log for clinical data would create a single entry for that data value. A log for debugging the network would record the request for the data, each of the likely multiple packets comprising the data transmission, and the acknowledgement message. Thus, debugging logs are a superset of, and take up much more space than, clinical data logs. A logging system should include options to allow users to select how detailed they want the logs to be. Data compression algorithms may be used to reduce the size of the log files provided that they do not lose data in the process.

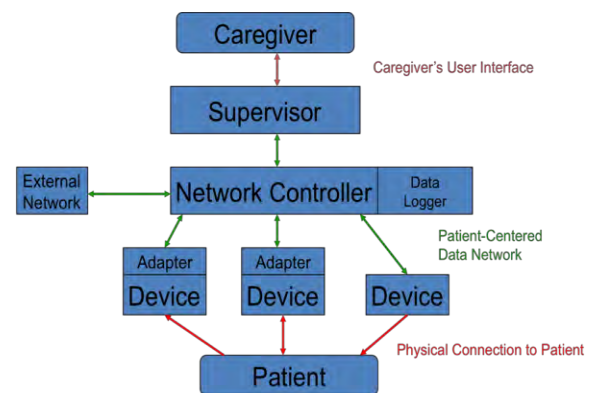


Figure 1. ASTM F2761-09 ICE Architecture Overview

Timestamps and Logical Clocks

The ICE network controller will contain a real-time clock set using the network time protocol (NTP). Synchronization of the network controller clock, and information about the accuracy with which it was set, will be entered in the log as it happens. Data from devices on the network will be entered in the log along with a sequence number (described below) and a timestamp from the network controller clock. The network controller will not attempt to set the device clocks or adjust the time they report, though some supervisor applications may adjust device clocks when possible. The data logger will record both network controller time (NTP time) and the times the individual devices report.

Often, the clock time of messages is not as

important as the sequence in which they are sent. Not all devices have clocks, and many devices that do have clocks only report the time to the nearest minute, or in the best case, second. This is too coarse-grained to properly order messages at the network controller level, and we cannot use the network controller clock to order them because messages may take varying times to travel through the network. This issue of ordering messages in a distributed system is a well-known problem in computer science, and the usual solution is to use logical clocks. Implementations such as Lamport clocks [7] or vector clocks [8] are applicable. We propose using vector clocks, where each device adapter on the network transmits a set of counter values with each transmission. This will allow analysis and playback programs to correctly establish causal ordering between messages even in cases where timestamps are not useful or available. Research results obtained through this analysis will inform an emerging Federal initiative on improving the timestamp accuracy of medical device data in the EHR.

Format of Device Data

Devices on the ICE network will transmit data using a standard format. ICE part 1 [2] does not specify this format, leaving its definition to future parts of the standard. It is expected that devices will encode their data using a well known ontology, though it is not necessary for all devices in the system to use the same ontology. Candidates include SNOMED, HL7, and I1073. One function of the playback and analysis software is to assist clinicians in categorizing adverse events. FDA CDRH uses event problem codes and evaluation codes to classify the device problems in associated with an adverse event. These codes are harmonized with ISO TS 19218 and there are plans to integrate these codes into SNOMED and to work with IEEE 11073 to incorporate the codes into these two codes sets to create a global vocabulary to report device problems.

We assume that the data logger playback application and supervisor applications will be able to interpret the ontologies used by connected devices. If this is not the case, the applications will at least be able to notify the user that the device is unsuitable for the application (in the case of the supervisor) or that the playback program cannot handle the

data log. The data logger will record raw network traffic even when it cannot interpret the contents.

Each data transmission from a device includes the unique device identifier (UDI) as specified by the FDA, a logical timestamp as described above, the data from the device encoded in that device's ontology of choice, and a checksum used to test if the data is corrupted in transmission. Where possible, existing adverse event ontologies will be used, such as the device problem and evaluation codes of the FDA's ★ MDR ★ system.

Security and Trustworthiness of the Log

When problems arise in systems whose components come from multiple manufacturers, it can be difficult to convince an individual manufacturer to take responsibility. The event recorder log provides a vendor-neutral record of transactions on the network that can be shown as evidence to device manufacturers.

We anticipate that the log from the recorder will be an important legal record as well as a clinical and engineering tool. This means that the data in the record must be trustworthy, and any tampering with the record must be obvious. To address these concerns, we give each log entry an individual sequence number and cryptographic signature in addition to a tagging it with the time the message was received at the network controller. The sequence number makes it obvious if a record is missing from the sequence, and the signature allows verification that the content of the record has not been changed.

Analysis and Replay of Log Data

An event log is only useful if it can provide relevant information to users. Turning the raw data in the log into useful information is the job of the replay program. This program should be able to open the data log from the event recorder, check it for consistency by examining the signature of each entry, and provide the user with a set of tools for analyzing the data. The log serves two general purposes: it will support analysis of adverse events involving multiple devices and it will allow system developers to view low-level data for debugging their applications. These purposes require different playback tools and techniques. We call the first use clinical log playback and the second use

debugging playback.

The clinical log playback tool will allow analysts to build an interactive time-line of logged data and events and to link text from clinician interviews in the appropriate places. Location information will be automatically included in the timeline when it is available in the record.

Analysts will need to be able to view the sequential data stream from a specific individual device and the interleaved sequences from multiple devices. In addition to the textual display, the program will be able to build a graphical timeline of data values and events from the devices. Because clinician narratives are an important part of the adverse event analysis process, the playback program will allow analysts to display narrative text beside the logged data, tag sections of the entries with times, and mark entries in the graphical timeline.

A typical session using the tool will have these steps:

1. Copy the log from the network controller to the computer running the playback tool.
2. Open the log in the tool and, for each device of interest, select the variables or items to appear on the graphical timeline. This step is illustrated in Figure 2.
3. Add clinician narratives from interviews. Manually mark the narrative with times given by the interviewed clinician. The tool will support vague times like “between 9 and 9:30 am” as well as descriptions like “between when the alarm went off the first and second time”.
4. The user can view a synchronized timeline of events and produce text or graphical output to help analyze the sequence of events and produce reports.

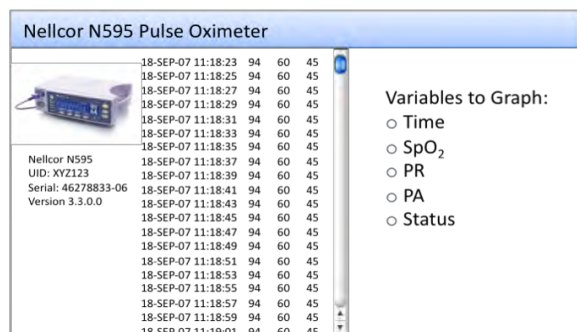


Figure 2. Device Data Selection User Interface

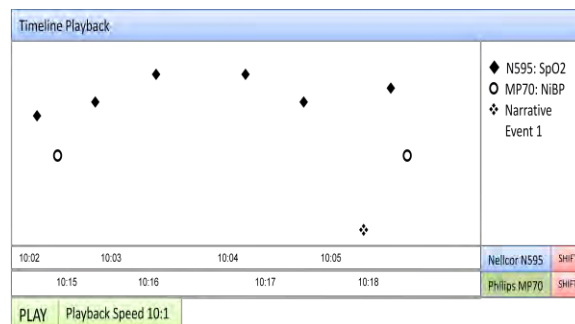


Figure 3. Mock-up of Clinical Playback User Interface

Figure 3 shows a mock-up of the clinical data playback user interface. The center of the display shows graphs of the device data that the user has selected and marks on the timeline for chosen events from narrative descriptions. There are two timelines along the bottom of the screen because the device data comes from two devices. The red “shift” button to the right of the timeline will allow the user to move the timelines forward and backward with respect to one another. The user can also let the system align events automatically using the clock and timestamp data in the log, but the manual option will be particularly useful for showing narrative events or events from manually entered or paper records. The user can also play back the data, either in real-time or at increased or reduced speeds.

Debugging playback typically involves too much data to view graphically. If system developers want to see a graphical display, they can use the clinical playback program or graph the data in another application. We expect that system developers will use standard tools like Matlab and protocol analyzer software to examine the files, and we will support this by exporting the data in appropriate formats. The playback tool for debugging will allow these users to select which data they want and pick an output format. The tool will also support down-sampling the data to reduce the size of the files and the strain on the analysis tools.

3 Future Work and Conclusion

Since no existing medical devices provide data at the resolution we want to support, we will add this capability to the Generic Infusion Pump [1], an open-source infusion pump project supported by the FDA. This prototype device will be useful for requirements gathering and for testing the

prototype system. We will connect existing devices in the MD PnP Interoperability Lab [9], including pulse oximeters, two models of Dräger ventilators, and a Philips patient monitor, although these legacy devices will not transmit high-resolution low-level data such as key presses.

Our data playback and visualization applications will be an improvement in current practice regardless of data source and even if the devices are not connected to an ICE network. A comprehensive log might make it possible to have greater contextual information to better understand the sequence of actions involved in an adverse event and hence more accurately and meaningfully report to FDA under the MDR regulations. The visualization and playback application will be useful with current hospital adverse event analysis workflow, although some data would have to be entered manually or converted from the devices proprietary formats.

Time synchronization and management of medical device clocks is becoming widely recognized as a barrier to acquiring accurately time-stamped EMR data for meaningful use and adoption of device data into EMRs. Without accurate timestamps, the information is of limited use for adverse event analysis.

We plan to explore integrating our data logger with the ASTER-D project. This project involves pulling patient history data and other relevant data from the EHR, applying event codes, and automatically transmitting an event report to the FDA. We may also be able to feed data into a Clinical Medical Device Management System.

We will produce general-purpose tools, but it will be useful for our development to focus on some concrete use cases. We will work with our MGH and FDA collaborators to identify appropriate and rich use cases, and our preliminary discussions have already identified unintended intra-operative awareness under anesthesia as an interesting case. This use case involves data from many different devices,

hand-written and computer entered case notes, and interviews with clinicians. When we are able to weave this data together into a coherent picture of what happened during a particular case, we will be well on our way to finishing the general purpose tools.

References

1. D. Arney, R. Jetley, P. Jones, I. Lee, and O. Sokolsky. Formal methods based development of a PCA infusion pump reference model: Generic infusion pump (GIP) project. In *Joint Workshop on High Confidence Medical Devices, Software, and Systems and Medical Device Plug-and-Play Interoperability* (HCMDSS-MDPnP 2007), 23-33, 2007.
2. ASTM F2761-09. ASTM F2761-09, New Specification for Equipment in the Integrated Clinical Environment – Part I: General Requirements for Integration.
3. S. L. Brown, R. A. Bright, and D. R. Tavriss. *Medical Device Epidemiology and Surveillance*. Wiley, 2007.
4. FDA Center for Devices and Radiological Health (CDRH). Workshop on medical device interoperability. http://www.mdnp.org/FDA_Interop_Workshop.php, January 2010.
5. Food and Drug Administration. Medical devices; medical device data systems, <http://www.federalregister.gov/articles/2011/02/15/2011-3321/medicaldevices-medical-device-data-systems>, February 2011.
6. ISO 80001-1:2010. *Application of risk management for IT-networks incorporating medical devices - Part 1: Roles, responsibilities and activities*. ISO, Geneva, Switzerland, 2010.
7. L. Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558-565, 1978.
8. F. Mattern. Virtual time and global states of distributed systems. In *Workshop on Parallel and Distributed Algorithms*, page 215-226, Chateau de Bonas, France, 1988. Elsevier.
9. MD PnP. Medical Device Plug and Play Program, <http://www.mdnp.org>.

Using Failure Modes, Mechanisms, and Effects Analysis in Medical Device Adverse Event Investigations

Shunfeng Cheng, Diganta Das, Michael Pecht

Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park, MD, USA

Abstract. In the United States, when medical devices are associated with adverse events that result in death or serious injury, or have malfunctions that could lead to death/serious injury, these events must be reported to the Food and Drug Administration's Center for Devices and Radiologic Health by device manufacturers and user facilities. However, the defects in the medical device evaluation process (e.g., failing to identify the failure mechanisms), can result in assessment risks and reoccurrences of adverse events. This paper presents an approach for medical device evaluation by using failure modes, mechanisms, and effects analysis to identify the root causes and failure mechanisms, which can improve the designs and reliability of medical devices. This method can also help medical device manufacturers to generate an internal evaluation reports for medical device evaluation, which can improve the reporting process to Food and Drug Administration.

Keywords: Medical device; adverse event; failure modes, mechanisms, and effects analysis.

1 Introduction

A medical device is an instrument, implant, or in-vitro reagent which is intended for use in the diagnosis of disease or other condition, or in the cure, mitigation, treatment, or prevention of disease, or to affect the structure or any function of the body, and which is not a drug or biologic product [1]. Manufacturers, user communities (e.g., hospitals or patients), and the Food and Drug Administration (FDA) all devote resources to ensure that medical devices are developed and used in a safe and effective manner throughout their lifetime. Even with this amount of oversight, devices still fail, resulting in adverse events, which are defined by FDA as “any undesirable experience associated with the use of a medical product.”

In the United States, adverse events related to medical devices are collected by FDA in Medical Device Reports (MDRs). Manufacturers conduct evaluation on the adverse event related medical device and report the evaluation results to FDA by some evaluation codes [2, 3], which are used to describe the methods, results, and conclusions following the evaluation of a device involved

in an adverse event. Evaluation method codes are used to indicate how the adverse event or failure was analyzed by the manufacturer, such as electrical or mechanical tests or visual examination. The outcome of this analysis is recorded using evaluation result codes, such as incomplete labeling. Finally, evaluation conclusion codes are used to summarize the manufacturer's findings of the analysis and focus on root causes to determine why the event occurred (for example, “device failure indirectly contributed to events”). Manufacturer evaluation codes are asked for in item H.6 in Form 3500A [4]; additional manufacturer narrative is asked for in item H10 to provide complementary information on the manufacturer evaluation.

Currently, there are no FDA guidance documents to guide manufacturers on the best practices for failure tracking and analysis. Device manufacturers might not use effective root cause analysis procedures leading to improper assignment of causes to the device failure. This may lead to the risk of reoccurrence of failures and inhibit the tracking of problems.

Center for Advanced Life Cycle Engineering (CALCE) at the University of Maryland have developed a failure modes,

mechanisms, and effects analysis (FMMEA) based approach that helps manufacturers improve their medical device evaluation processes and the future products by providing a systematic evaluation method for potential failures and causes, making it more efficient to identify the root causes and mechanisms of the failure of devices.

2 Using FMMEA for Adverse Event Related Medical Device Evaluation

FMMEA (Figure 1) is a means to help manufacturers implement medical device evaluation in a systemic manner that allows for investigation of the failure mechanisms and generate manufactures' internal device-specific evaluation reports in a failure site-mode-cause-mechanism structure, which may aid in reporting adverse event related medical device evaluations to FDA.

No literature or reports have shown that medical device manufacturers are using FMMEA, although some similar abbreviations were reported. For example, failure modes and effects analysis (FMEA) or failure modes, effects, and criticality analysis (FMECA) are used to identify the possible failure modes and causes of medical devices [5]. FMEA or FMECA methodologies outline procedures to recognize and evaluate the potential failure of a product and its effects and to identify actions that could eliminate or reduce the likelihood of the potential failure to occur [6]. Many organizations within the electronics industry have employed or required the use of FMEA, but in general this methodology has not provided satisfaction, except for the purpose of safety analysis [7]. A limitation of the FMEA methodology is that it does not identify the product failure mechanisms in the analysis and reporting process. For example, when conducting FMEA on infusion pumps [5], the important failure mechanisms for catheter system leakage or breakage, such as fatigue, corrosion, kink, and chemical precipitation, were not identified, and as a result the design update process would not target those mechanisms.

FMMEA is a tool to support physics-of-failure based design for reliability [7-10]. It can identify potential failure mechanisms for

all potential failures modes and prioritize the failure mechanisms. FMMEA can aid medical device manufacturers in the development of reliable designs, planning tests, and screens to validate nominal design and manufacturing specifications and determine the limits on the level of defects introduced by the variability in manufacturing and materials. FMMEA enhances the value of traditional FMEA methodologies by identifying the high-priority failure mechanisms in order to create an action plan to mitigate their effects.

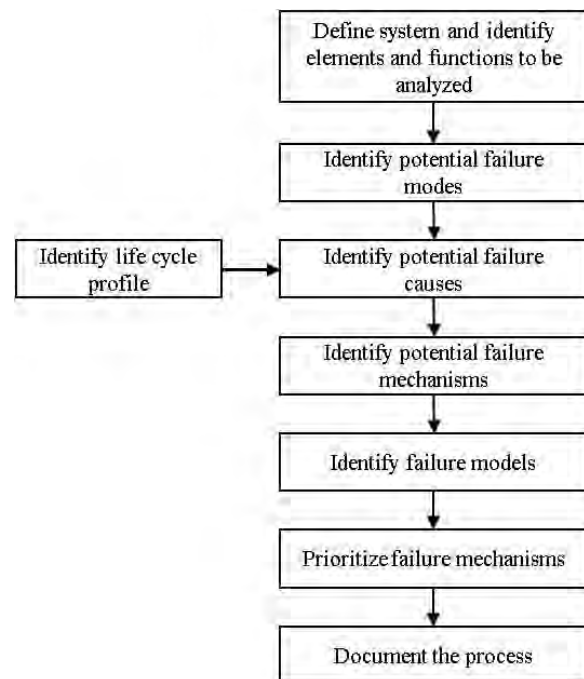


Figure 1. FMMEA Methodology [8-10]

FMMEA uses the life cycle profile (LCP) of a product along with the design information to identify the critical failure mechanisms affecting a product. An LCP is a forecast of the events and the associated environmental and usage conditions a product may experience from manufacture to end of life. The device is divided into its lower level subassemblies for investigation. For medical devices, FMMEA can be conducted down to the lowest level at which the device manufacturer still has design control; FMMEA at lower levels should be performed by subsystem or component vendors. These subassemblies are potential sites of failure. In FMMEA the potential failure modes for each failure site are listed. A failure mode is the manner in which a failure

is observed by methods such as visual inspection, electrical measurement, or other tests and measurements. For each failure mode, the potential failure causes are analyzed. A failure cause is the specific process, design, and/or environmental condition that initiate a failure and whose removal will eliminate the failure. Possible failure causes are investigated in the entire life cycle of the device, including design, manufacture, operation, and maintenance. For example, in a multilayer ceramic capacitor (MLCC), a component used in medical devices [11], the failure modes may be short, open, or parameter shift, such as a decrease in insulation resistance or an increase in dissipation factor. The potential causes of these failures may be operational temperature and humidity conditions during storage or transportation.

Next, potential failure mechanisms are identified. Failure mechanisms are the processes by which a specific combination of physical, electrical, chemical, biological, and mechanical stresses induces failures. Using MLCCs under temperature-humidity-bias conditions as an example, the dominant failure mechanisms include metal migration between the electrodes, dielectric degradation caused by moisture penetrating the voids, and creation of oxygen vacancies in the dielectric of the capacitor.

During the life cycle of a product, several failure mechanisms may be activated by different environmental and operational parameters acting at various stress levels, though, in general, only a few operational and environmental parameters and failure mechanisms are responsible for the majority of failures. In the process of conducting FMMEA, we assess the combinations of occurrence and severity of each failure mechanism, where the probability of occurrence is taken into consideration from the distributions of the loads and the geometric/material features, while the severity is obtained from the seriousness of the effects of the failure caused by a particular mechanism.

Medical device manufacturers can conduct FMMEA internally to identify the potential failure sites, modes, causes, and mechanisms of a medical device. The use of FMMEA will enable manufacturers to create an internal evaluation report organized in the failure site-mode-cause-mechanism structure, as shown in Figure 2. Another benefit of conducting FMMEA is that it would help manufacturers monitor and improve the reliability of their products and provide manufacturers with useful information to investigate and correct the adverse events.

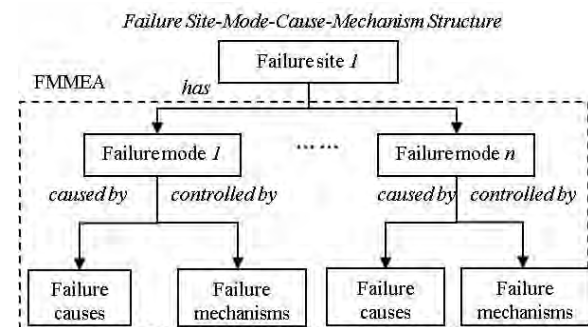


Figure 2. Failure Site-Mode-Cause-Mechanism Structure for Adverse Event Investigation

3 Example: FMMEA on Infusion Pump Failure

An external infusion pump is used to deliver fluids into a patient's body in a controlled manner. FDA has seen an increase in the number and severity of infusion pump related adverse events [12]. An example of FMMEA of the flow generation and regulation system of an infusion pump is shown in Table 1.

Generally, the infusion pump contains three main subsystems: the fluid reservoir, a catheter system for transferring fluids into the body, and a flow generation and regulation system that combines electronics (e.g., processor, memory, and power management module) with a flow control mechanism (e.g., pump and sensors) to generate and regulate flow [13].

Potential failure sites		Potential failure modes	Potential failure causes	Potential effects*	Potential failure mechanisms
1 Flow generation and regulation system	Power (only battery is concerned)	Voltage error, unable to be charged, overheating	Battery depleted, overcharged, degraded	Underdose, overdose, therapy delay	Battery wear-out
	Pump	Pumps inaccurate size/rate of dose (including “fail to pump”), operating abnormally	Component defects; improper position of pump; failure to release inside air, lower inside air pressure; ambient temperature, humidity, air pressure too high or low; design error; labeling error; insufficient training; calibrating or programming error	Underdose, overdose, therapy delay, free flow, air in line, reverse flow	Wear-out, fatigue, corrosion
	Control module: software	Runtime error, incorrect messages, false alarms, failure to alarm, incorrect dose calculation	Buffer overflow or underflow; incorrect dynamic libraries; uninitialized variables; wrong algorithms or programming, threshold setting error; insufficient training	Underdose, overdose, or therapy delay	Design errors
	Control module: hardware (e.g., processor, memory)	Overheating, short or open circuit, high leakage current, high or low impedance, missed alarm, false alarm, fail to read/write data	Insufficient cooling, shielding or insulation; non-human interference; loose interconnection; corrosive fluid ingress; component failure, sensor contaminated, out of calibration; design error; labeling error; insufficient training.	Underdose or overdose, electric shock, therapy delay, contamination	Overstress or wear-out, fatigue, corrosion, radiation
	User interfaces (e.g., display)	Cracks in package or case, broken keypad, key stuck /depressed, speaker/audio unit failure	Incorrect operation, environmental effects, accidents (e.g., falling), fluid ingress, design defects, component defects, component degraded; design errors; labeling errors; insufficient training	Under-dose or over-dose, contamination, therapy delay	Wear-out, overstress, corrosion, fatigue, radiation, creep

Table 1. Examples of FMMEA on Infusion Pumps (Excluding biological or chemical hazards or failures)

*Information from this column is used just for determination of severity and prioritizing the critical mechanisms.

When an adverse event related to an infusion pump is reported, the manufacturer can identify potential failure sites and modes based on the description of the adverse event. Manufacturers then refer back to FMMEA evaluation results to find the possible causes and mechanisms, and then conduct actual inspection to validate the failure sites, root

causes, and mechanisms, and then have an internal report about the evaluation results. For example, if the device problem was reported as “failure to alarm”, which is failure mode, and patient problem code was “over-dose”, which is failure effect, the potential failure sites may include the software and related components. If the failure site was

confirmed as “control module: hardware”, the potential failure causes and mechanisms could be determined. The final evaluation could be reported as shown in Figure 3.

4 Discussions and Conclusions

FMMEA enables manufacturers to narrow down device failures to a desired level of abstraction (system, subsystem, or component), identify the root causes and mechanisms of the failures, take proper actions to reduce the recurrence of the failures, and improve device design, product realization, and sustainment. If the manufacturer has a family of similar medical devices that may be used in similar environmental and operational conditions, FMMEA evaluation results could be transferred to other devices in the family. With more root causes of device failure have been identified and controlled, medical devices can be expected to have better reliability. This can reduce the number of medical device–related adverse events. Manufacturer can utilize knowledge of a product’s life cycle loading and failure mechanisms and models identified by FMMEA to assess reliability of medical devices. The possible failures of a medical device can be cataloged by FMMEA, and potential risks can continue to be updated by monitoring the device’s life cycle environmental and usage conditions while taking into consideration the devices geometry

and material properties. Adverse event possibilities can then be identified and averted based on that knowledge.

We are working with computer scientists within FDA to determine what the data structure might look like. However, we do not want prescribe for manufacturers the particular FMMEA data format that they will integrate into their design process. Manufacturers will store data in a format that is compatible with and accessible to their adverse event resolution process.

When reporting to FDA, manufacturers need not send the complete FMMEA evaluation to FDA, but share the parts related to a specific adverse event. When an adverse event is reported, the manufacturer could use existing FMMEA to narrow in on the potential failure modes and causes and report to FDA using the linked evaluation codes after actual validation. A failure site-mode-cause-mechanisms structure with explanations can provide content rich information in the text fields when reporting MDRs. One effect of FMMEA on adverse event ontology is that FMMEA can generate new medical device-specific codes. The device-specific codes used in failure-site-mode-mechanism structured reports can be used to extend current adverse event ontology beyond the current generic reporting terminology. However, the evolution of adverse event ontology in a general sense is beyond the scope of this paper.

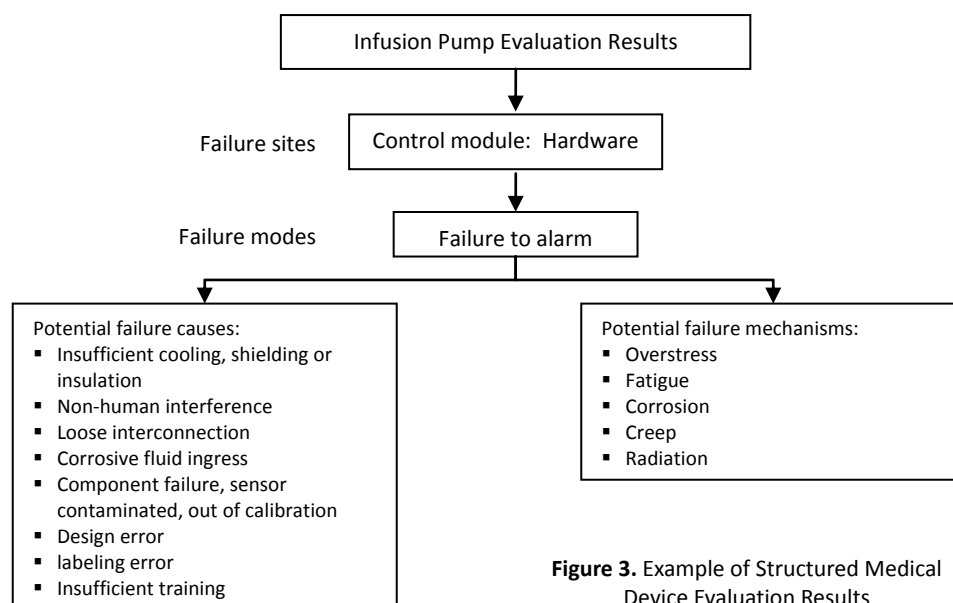


Figure 3. Example of Structured Medical Device Evaluation Results

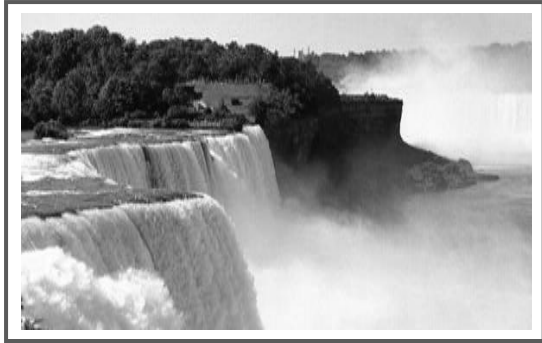
Acknowledgment

The authors would like to thank the companies and organizations that support research activities at the Center for Advanced Life Cycle Engineering at the University of Maryland. The authors would also like to thank Dr. Sandy Weininger and Dr. Raoul Jetley from FDA for their guidance and help.

References

1. Food and Drug Administration, "Is the Product a Medical Device?", <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ClassifyYourDevice/ucm051512.htm>, accessed on May, 3, 2011.
2. Food and Drug Administration, "Coding Concepts," <http://www.fda.gov/MedicalDevices/Safety/ReportaProblem/EventProblemCodes/ucm134747.htm>, accessed on May, 3, 2011.
3. Food and Drug Administration, "Appendix - B: Device and Patient Problem Codes, Manufacturer Evaluation Method, Manufacturer Evaluation Results, and Manufacturer Conclusion Codes," <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm106742.htm>, accessed on May, 3, 2011.
4. Food and Drug Administration, "Instructions for Completing Form FDA 3500A," <http://www.fda.gov/Safety/MedWatch/HowToReport/DownloadForms/ucm149238.htm>, accessed May, 3, 2011.
5. Fechter, R.J., and Barba J.J., "Failure Mode Effect Analysis Applied to the Use of Infusion Pumps," Proceedings of the 26th Annual International Conference of the IEEE EMBS, pp. 3496-3499, San Francisco, CA, 2004.
6. SAE Standard. SAE J1739: Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) and Effects Analysis for Machinery (Machinery FMEA). Warrendale, PA: SAE. 2002.
7. Ganesan, S., Eveloy, V., Das, D., and Pecht, M. Identification and utilization of failure mechanisms to enhance FMEA and FMECA. Proc IEEE Workshop on Accelerated Stress Testing & Reliability (ASTR). 2005.
8. Pecht, M., *Prognostics and Health Management of Electronics*. Wiley-Interscience, New York, NY, 2008.
9. Pecht, M. and Dasgupta, A., "Physics-of-Failure: An approach to reliable product development," *Journal of the Institute of Environmental Sciences*. 1995: 38(5), pp.30-34.
10. Pecht, M. and Gu, J., "Physics-of-failure-based Prognostics for Electronic Products," *Transactions of the Institute of Measurement and Control*, Vol. 31, No. 3/4, 2009, pp. 309-322.
11. National Institution of Standards and Technology, "Medical Device Reliability," http://www.nist.gov/msel/materials_reliability/cell_tissue_mechanics/Medical-Device-Reliability.cfm, accessed May, 3, 2011.
12. Food and Drug Administration, "White Paper: Infusion Pump Improvement Initiative," <http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/GeneralHospitalDevicesandSupplies/InfusionPumps/ucm205424.htm>, accessed May, 11, 2010.
13. Texas Instruments, "Infusion Pump Block Diagram," <http://focus.ti.com/docs/solution/folders/print/481.html>, accessed on May, 11, 2010.

Working with Multiple Biomedical Ontologies



ICBO

International Conference on Biomedical Ontology

July 26, 2011
Buffalo, New York, USA

NIFSTD and NeuroLex: A Comprehensive Neuroscience Ontology Development Based on Multiple Biomedical Ontologies and Community Involvement

Fahim T. Imam, Stephen D. Larson, Jeffery S. Grethe, Amarnath Gupta,
Anita Bandrowski, Maryann E. Martone

Neuroscience Information Framework, Center for Research in Biological Systems,
University of California, San Diego, La Jolla, USA

Abstract. A core component of the Neuroscience Information Framework (NIF) project, the NIF Standard (NIFSTD) ontology, was envisioned as a set of modular ontologies that provide a comprehensive collection of terminologies to describe neuroscience relevant data and resources. We present here on the structure, design principles, community engagement and current state of NIFSTD that reuses multiple biomedical ontologies, applied for an effective semantic search mechanism.

Keywords: Multiple ontologies, modularity, ontology reuse, neuroscience ontology, semantic search

1 Introduction

The Neuroscience Information Framework (NIF)¹ project involves discovering and utilization of various neuroscience resources over the web. The end product is a semantic search engine and discovery portal consisting of a framework that describes resources and provides simultaneous access to multiple types of information, organized by various relevant categories.

Behind the scenes, NIFSTD [1] is a critical constituent for the NIF project in order to enable its effective concept-based search mechanism against a diverse collection of web based neuroscience data and resources. The overall ontology is assembled in a form that promotes reuse of multiple ontologies, easy extension and modification over the course of its evolution. NIFSTD relies on existing biomedical ontologies as the initial building blocks. These ontologies currently include CHEBI, Gene Ontology (GO), Protein Ontology (PRO), Ontology for Biomedical Investigations (OBI), and The Ontology of Phenotypic Qualities (PATO). Section 2 of this paper describes the basic structure and design

principles of NIFSTD. The relation between NIFSTD and the NeuroLex wiki environment for collaborative development is the focus of Section 3. Current state and progress of NIFSTD is highlighted in Section 4 followed by conclusion and future work in Section 5.

2 NIFSTD Basic Structure and Design Principles

NIFSTD follows OBO Foundry [11] principles as long as they are reasonable for NIF's application purposes. While the principles promote developing highly interoperable and reusable reference ontologies in ideal cases, following some of them in a rigid manner has often proven to be too ambitious for day-to-day development. Some of the principles may become impractical for extending existing ontology modules especially when there is a deadline constraint imposed for those extensions. Gaining OBO Foundry community consensus for a production system is difficult as we often need to move quickly along with the project. Therefore, we rather favor a system whereby we start with minimal complexity as required and add more as the ontologies evolve over time towards perfection.

¹ The Neuroscience Information Framework (NIF),
<http://neuinfo.org>

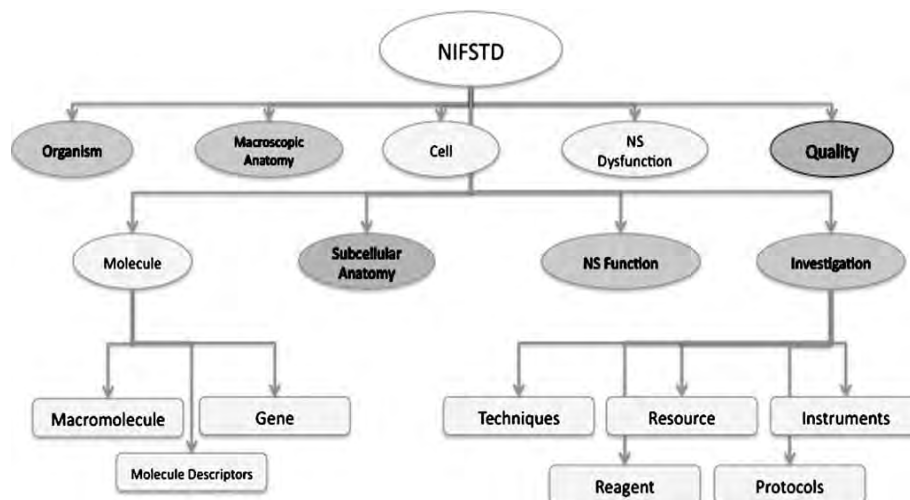


Figure 1. The semantic domains covered in the NIFSTD ontology. Separate OWL modules cover the domains specified within the ovals. The umbrella file <http://purl.org/nif/ontology/nif.owl> imports each of these modules when opened in Protégé. Each of the modules, in turn, may cover multiple sub-domains, some of which are shown in the rectangular boxes.

Domain	External Source	Import/ Adapt	Module
Organism taxonomy	NCBI Taxonomy, GBIF, ITIS, IMSR, Jackson Labs mouse catalog	Adapt	NIF-Organism
Molecules	IUPHAR ion channels and receptors, Sequence Ontology (SO), ChEBI, and Protein Ontology (PRO); pending: NCBI Entrez Protein, NCBI RefSeq, NCBI Homologene, NIDA drug lists	Adapt IUPHAR, ChEBI; Import PRO, SO	NIF-Molecule NIF-Chemical
Sub-cellular	Sub-cellular Anatomy Ontology (SAO). Extracted cell parts and subcellular structures. Imported GO Cellular Component with mapping	Import	NIF-Subcellular
Cell	CCDB, NeuronDB, NeuroMorpho.org. Terminologies; pending: OBO Cell Ontology	Adapt	NIF-Cell
Gross Anatomy	NeuroNames extended by including terms from BIRN, SumsDB, BrainMap.org, etc; multi-scale representation of Nervous System Mac Macroscopic anatomy	Adapt	NIF-GrossAnatomy
Nervous system function	Sensory, Behavior, Cognition terms from NIF, BIRN, BrainMap.org, MeSH, and UMLS	Adapt	NIF-Function
Nervous system dysfunction	Nervous system disease from MeSH, NINDS terminology; Disease Ontology (DO)	Adapt/Import	NIF- Dysfunction
Phenotypic qualities	PATO Imported as part of the OBO foundry core	Import	NIF-Quality
Investigation: reagents	Overlaps with molecules above, especially RefSeq for mRNA	Import	NIF-Investigation
Investigation: instruments, protocols, plans	Based on Ontology for Biomedical Investigation (OBI) to include entities for biomaterial transformations, assays, data collection, data transformations.	Adapt	NIF-Investigation
Investigation: resource type	NIF, OBI, NITRC, Biomedical Resource Ontology (BRO)	Adapt	NIF-Resource
Biological Process	Gene Ontology's (GO) biological process in whole	Import	NIF-BioProcess
Cognitive Paradigm	Cognitive Paradigm Ontology (CogPO)	Import	NIF-Investigation

Table 1. Domains covered by the current NIFSTD along with the vocabularies imported from the external, community sources and the corresponding OWL modules.

Modularity. One of the key features of NIFSTD is that its ontologies are built in a modular fashion, each module covering orthogonal domain of Neuroscience concepts (Fig.1). These distinct domains include macroscopic anatomy, cell types, techniques, nervous system function, molecules etc. Following a powerful modularization ontology design pattern [3], NIFSTD promotes easy extendibility towards its evolution. It avoids duplication of efforts by conforming to standards that promote reuse. Each of the modules is standardized to the same upper level ontologies such as the Basic Formal Ontology (BFO) and OBO Relations Ontology (OBO-RO).

Representation Language. Expressed in W3C standard OWL Description Logic (OWL-DL) dialect, NIFSTD ensures computational decidability and supports automated reasoning via common DL reasoners such as Pellet and FACT++. The reasoners ensure the ontologies to be kept in a logically consistent state and allow computing inferred classification of asserted class hierarchies.

Reuse of Available Ontology Sources. Wherever possible, NIFSTD reuses existing available distilled knowledge sources, terminologies and ontologies. Community sources were culled from a variety of sources, ranging from fully structured ontologies to loosely structured controlled vocabularies. Table 1 illustrates various domains in NIFSTD that are either adapted or imported as a whole from various external, community sources. The process of importing or adapting a new ontology or vocabulary source varies depending upon its state [1].

- If a source already uses OWL, the OBO-RO and the BFO and is orthogonal to existing modules, the import simply involves adding an owl:import statement to the main ontology file (nif.owl).
- If an existing orthogonal ontology is in OWL but does not use the same foundational ontologies as NIFSTD, then an ontology-bridging module (explained later) is constructed declaring the deep level semantic equivalencies such as foundational objects and processes.

- If an external source is satisfiable by the above two rules but observed to be too large for NIF's scope of interests, a relevant subset is extracted as suggested by NIF's domain experts (e.g., CheBI, OBI etc. are adapted rather than imported as a whole). During our recent extensions, we have been following MIREOT principle [6] that allows extracting a required portion of a larger ontology.
- If the external source has not been represented in OWL, or does not use the same foundation as NIFSTD, then the terminology is adapted to OWL/RDF in the context of the NIFSTD foundational layer ontologies.

Single Inheritance. NIFSTD follows the simple inheritance principle for the hierarchy of *named* classes; i.e., an asserted *named* class can have only one named class as its superclass; however, a named class can have multiple *anonymous* superclasses. This principle promotes the named classes to be univocal and to avoid ambiguities. The classes with multiple superclass are derivable via automated classification on defined NIFSTD classes with necessary and sufficient conditions. This approach saves a great deal of manual labor and minimizes human errors inherent in maintaining multiple hierarchies. Also, this approach provides logical and intuitive reason as to how a class may exist in multiple, different hierarchies. Motivations behind this approach can be found in detail in [10].

Unique Identifiers and Annotation Properties. NIFSTD entities are identified by a unique identifier and accompanied by a variety of annotation properties derived from Dublin Core Metadata (DC) and the Simple Knowledge Organization System (SKOS) model. We reuse the same URI through MIREOTed extracted classes from the source, which allows us to avoid extra mapping annotations with the community sources, e.g., GO or PRO class identifiers remain unaltered. A tool based on MIREOT principles called OntoFOX [4] is proven to be useful in this regard.

Object Properties and Bridge Modules. NIFSTD object properties are mostly drawn from the OBO Relations Ontology (OBO-RO).

Fig. 2 illustrates some example object properties between various NIFSTD Classes.

The cross-domain relations are specified in separate *bridging* modules – modules that only contain logical restrictions and definitions on a required set of classes assigned between multiple modules [1]. The bridging modules allow the main domain modules – e.g., anatomy, cell type, etc. to remain independent of one another without the bridging modules (e.g., Fig. 3). They help to keep the modularity principles intact, and facilitate extensions for broader communities without worries about specific, NIF-centric views on how a class should be logically

defined or restricted. These bridging modules can easily be excluded in order to focus on core modules and build separate bridging module that are appropriate for user specific domain.

Versioning. Various annotation properties are associated with versioning different levels of content within NIFSTD. These include creation and modification dates for each of the classes; file level versioning for each of the modules, annotations for retiring antiquated concept definitions, tracking former ontology graph position and replacement concepts. Refer to [1] for more details on NIFSTD versioning policies.

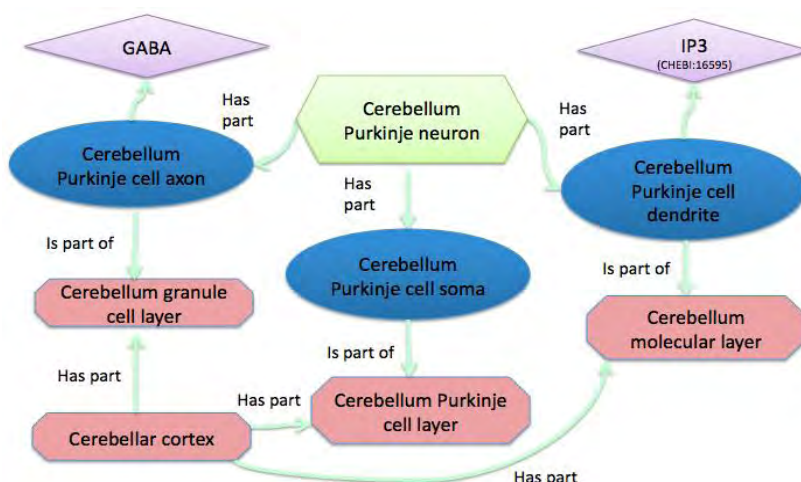


Figure 2. A typical knowledge model in NIFSTD. Both cross-modular and intra-modular classes are associated through object properties mostly drawn from the OBO Relations ontology (RO).

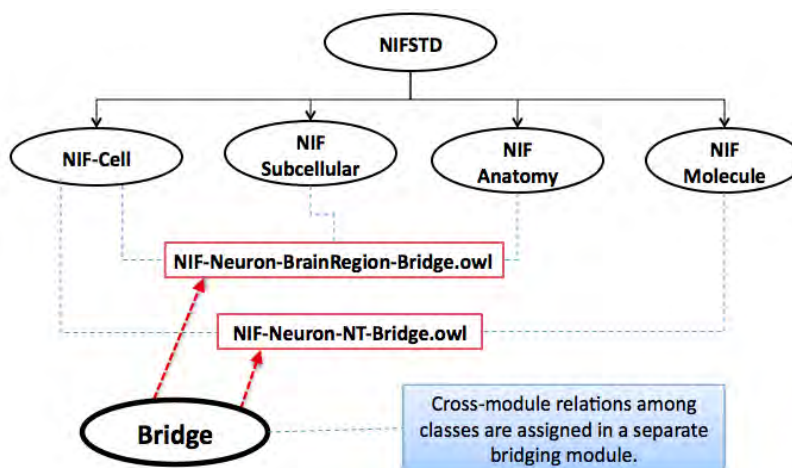


Figure 3. Two example bridging OWL modules in NIFSTD (rectangular boxes) that contain class property associations between multiple core modules.

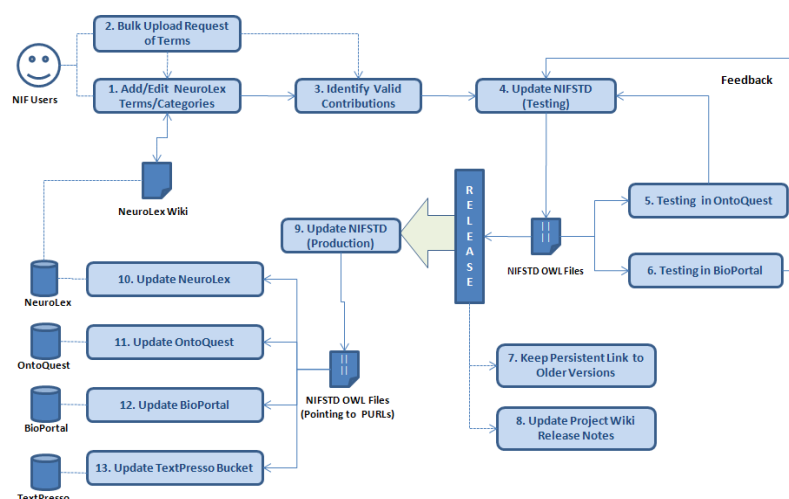


Figure 4. NIF workflow to transition knowledge between NeuroLex and NIFSTD.

Accessing NIFSTD Ontologies.² NIFSTD is primarily available in OWL format and can be loaded through Protégé [7] or similar OWL editing tools. It is also viewable through NCBO BioPortal [12]. Within NIF, NIFSTD is served through an ontology management system called OntoQuest [2]. OntoQuest generates an OWL-compliant relational schema and implements various ontological graph search algorithms for navigating, term searching and expanding query terms based on their logical restrictions for the NIF search portal. It also provides web services [8] to extract ontology contents. NIFSTD is also available in RDF and has a SPARQL endpoint [9].

3 NIFSTD and NeuroLex Wiki

One of the largest roadblocks that we encountered during our ontology development was the lack of tools for domain experts to view, edit and contribute their knowledge to NIFSTD. Existing editing tools were difficult to use or required expert knowledge to employ. We strive to balance the involvement of the neuroscience community for domain expertise and the knowledge engineering community for ontology expertise when constructing the NIFSTD. By combining open-source technologies related to semantic media wiki, NIF developed NeuroLex³, a semantic wiki environment to promote easy

collaboration between the neuroscience community and domain experts.

The NeuroLex Wiki. We envision NeuroLex as the main entry point for the broader community to access, annotate, edit and enhance the core NIFSTD content. The peer-reviewed contributions in the media wiki are later implanted in formal OWL modules. It should be noted that NIF is not charged with development of new modules but relies on community for new contents. Therefore, the NeuroLex wiki has proven to be ideal for NIF's current scope. For example, it has proven to be effective and helpful in the area of neuronal cell types where NIF is working with a group of neuroscientists, to create a comprehensive list of neurons and their properties.

NeuroLex Wiki Facts. NeuroLex provides a bottom-up ontology development approach where multiple participants can edit the ontology instantly. Control of content is done after edits are made based on the merit of the content, rather than by the blessing of a few known individuals. Semantics is limited to what is convenient for the domain. Essentially, the NeuroLex approach is not a replacement for top-down construction, but critical to increase accessibility for non-ontologist domain experts.

NeuroLex has become critically important for the large corpus domain with no formal categories where the entities are unstable and unrestricted with no clear edges. NeuroLex is potentially necessary when participants are uncoordinated users, amateur users, or naive

² The NIF Standard Ontologies, <http://ontology.neuinfo.org>

³ The NeuroLex Wiki Site, <http://neurolex.org>

cataloguers. NeuroLex provides various simple forms for structured knowledge where communities can contribute and verify their knowledge with ease. It also allows to generate a specific class hierarchy, or extraction of a specific portion of the ontology based on certain properties in a spreadsheet, without having to learn complicated ontology tools. The NIFSTD/NeuroLex development and curation workflow are depicted in Fig. 4.

NIFSTD vs. NeuroLex Properties. While the properties in NeuroLex are meant for easier interpretation, the restrictions in NIFSTD are usually based on rigorous OBO-RO standard relations. For example, the property ‘soma located in’ is translated as ‘Neuron X’ has_part some (‘Soma’ and (part_of some ‘Brain region Y’)) in NIFSTD. Sometimes we use similar kind of ‘macro’ relation such as ‘has_neurotransmitter’ within NIFSTD, recognizing that these relations can be specified more rigorously. These ‘macro’ relations readily lend themselves into rigorous

representations should they become necessary at a later date.

4 Current State and Progress

We use NIFSTD ontologies with the sense of being essentially useful for our semantic search interface. One of the most powerful features of having ontology is that it allows explicit knowledge of a domain to be *asserted* from which implicit logical consequences can be *inferred* using logical reasoners. The following example illustrates the strengths and usefulness of this feature for NIF search portal.

NIFSTD includes various neuron types with an asserted simple hierarchy within the NIF-Cell module (Fig. 5 is an example with five neuron types). However, we assert various logical necessary restrictions about these neurons in a bridging module where we also specify various *defined* neuron types with necessary and sufficient conditions as illustrated in Fig. 6.

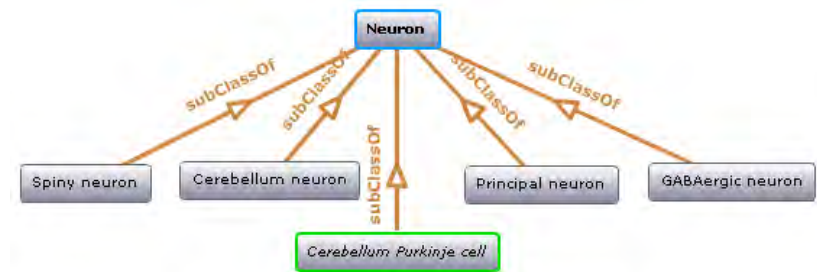


Figure 5. NIFSTD Cerebellum Purkinje cell is simply a subclass of a Neuron before invoking a reasoner along with asserted restrictions as specified in Fig. 5.

Class name	Asserted defining (necessary & sufficient) expression
Cerebellum neuron	Is a ‘Neuron’ whose soma lies in any part of the ‘Cerebellum’ or ‘Cerebellar cortex’
Principal neuron	Is a ‘Neuron’ which has ‘Projection neuron role’, i.e., a neuron whose axon projects out of the brain region in which its soma lies
GABAergic neuron	Is a ‘Neuron’ that uses ‘GABA’ as a neurotransmitter

Class name	Asserted necessary conditions
Cerebellum Purkinje cell	1. Is a ‘Neuron’ 2. Its soma lies within ‘Purkinje cell layer of cerebellar cortex’ 3. It has ‘Projection neuron role’ 4. It has ‘GABA’ as a neurotransmitter 5. It has ‘Spiny dendrite quality’

Figure 6. Typical NIFSTD asserted restrictions for various neuron types. The first table in the figure defines three neuron types with logical necessary and sufficient conditions. The second table lists a set of necessary restrictions for Cerebellum Purkinje cell. All these restrictions written in a readable format here is expressed in OWL DL language in actual NIFSTD.

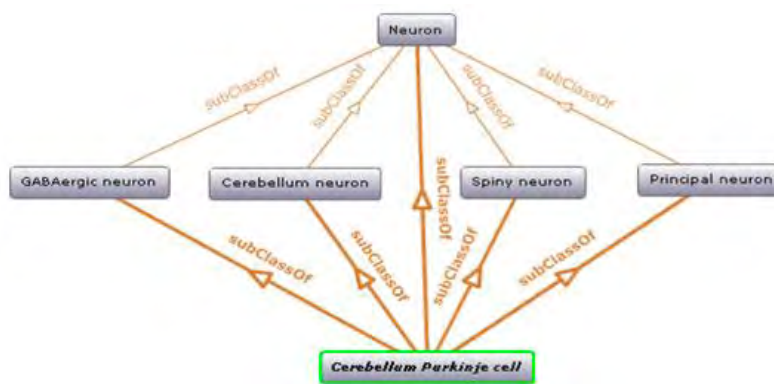


Figure 7. After invoking a reasoner NIFSTD Cerebellum Purkinje cell becomes a subclass of four different defined neuron types based on the restrictions specified in Figure 6.

When the NIF-Cell module along with the bridging modules is passed to a reasoner, the reasoner automatically computes for the asserted neuron types with restrictions (as indicated in Fig. 6) and produce a hierarchy where a neuron can have multiple inferred superclasses. In this example, although we did not explicitly state that Cerebellum Purkinje cell is anything other than a simple neuron, the reasoner identified that the neuron is an inferred subclass of four different defined neurons (Fig. 7) namely, GABAergic neuron, Cerebellum neuron, Spiny neuron and Principal neuron, based on the logical restrictions specified as in Fig. 6. Having the defined classes has enabled us to develop useful concept-based queries through the NIF search interface. For example, while searching for 'GABAergic neuron', the system recognizes the term as 'defined' from the ontology, and looks for any neuron that has GABA as a neurotransmitter (instead of the lexical match of the search term) and enhances the query over this inferred list of neurons. Searching these defined terms in a Google search would essentially exclude all the GABAergic neurons unless they are explicitly listed within the search.

The key feature of the current NIFSTD is the inclusion and enrichment of various cross-domain bridging modules. These modules contain necessary restrictions along with a set of defined classes to infer useful classification of neurons and molecules. A running list of defined concepts along with textual definitions can be found on the NIFSTD wiki page in [5]. The following list illustrates some of the

defined concepts in NIFSTD and their classification schemes:

- Neurons by their soma location in different brain regions – e.g., Hippocampal neuron, Cerebellum neuron, etc.
- Neurons by their neurotransmitter e.g., GABAergic neuron, Glutamatergic neuron, Cholinergic neuron
- Neurons by their circuit roles – e.g., Intrinsic neuron, Projection neuron
- Neurons by their morphology – e.g., Spiny neuron
- Neurons by their molecular constituents – e.g., Pervalbumin neuron, Calretinin neuron
- Classification of molecules and chemicals by their molecular roles – e.g., Drug of abuse, Neurotransmitter

The modularity principles along with the bridging modules allowed us to limit the complexity of the base ontologies and promoted easy extendibility. Also, it allowed us to rely on module-by-module reasoning for consistency checking, inferred subsumption and other reasoning tasks. We provide different inferred axioms saved in separate modules for the end-users convenience. We are still looking for a practical reasoning mechanism with more powerful inference engines that can scale with large ontologies like NIFSTD as a whole with all its individual modules imported together.

5 Conclusions and Future Work

The NIF project provides an example of practical ontology development and how it can be used to enhance search and data integration across diverse resources. Using the upper level BFO ontologies allowed us to promote a broad semantic interoperability between a large numbers of biomedical ontologies. We have defined a process to form complex semantics to various neuroscience concepts through NIFSTD and through NeuroLex collaborative environment. NIF encourages the use of community ontologies for resource providers, and as the project moves forward, we are using NIFSTD to build an increasingly rich knowledgebase for neuroscience that integrates with the larger life science community.

Acknowledgement

Supported by a contract from the NIH Neuroscience Blueprint HHSN271200800035C via NIDA.

References

1. Bug, W.J., Ascoli, G.A., Grethe, J.S., Gupta, A., et al.: The NIFSTD and BIRNLex Vocabularies: Building Comprehensive Ontologies for Neuroscience. *Neuroinformatics*. 2008 Sep;6(3):175-94. PMID: 18975148 (2008)
2. Gupta, A., Bug, W., Marengo, L., Condit, C., et al.: Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF). *Neuro-informatics*. 2008 Sep;6(3):205-17. PMID: 18958629 (2008)
3. Ontology Design Patterns (ODPs) Public Catalog, <http://odps.sourceforge.net>
4. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. *BMC Research Notes*. 2010, 3:175. PMID: 20569493 (2010)
5. NIFSTD Defined Concept List, <http://ontology.neuinfo.org/defined-types.html>
6. Courtot, M., Gibson, F., Lister, A., Malone, J., Schober, D., Brinkman, R., Ruttenberg, A.: MIREOT: the Minimum Information to Reference an External Ontology Term. Available from *Nature Precedings*, <http://dx.doi.org/10.1038/npre.2009.3576.1>, (2009)
7. Protégé Ontology Editor and Knowledge Acquisition System, <http://protege.stanford.edu>
8. OntoQuest Web Services, <http://ontology.neuinfo.org/ontoquest-service.html>
9. NIFSTD SPARQL Endpoint, <http://ontology.neuinfo.org/sparql-endpoint.html>
10. Rector, A.: Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. *Proc K-CAP* (2003)
11. Smith, B., Ashburner, M., Rosse, C., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*. 1251-1255 (2007)
12. NIFSTD in NCBO BioPortal, <http://bioportal.bioontology.org/ontologies/1084>

Use of Multiple Ontologies to Characterize the Bioactivity of Small Molecules

Ying Yan¹, Janna Hastings¹, Jee-Hyub Kim¹, Stefan Schulz²,
Christoph Steinbeck¹, Dietrich Rebholz-Schuhmann¹

¹European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

²Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

Abstract. ChEBI is an ontology of biologically interesting chemicals. Biological activities of chemical entities comprise interactions with biological entities such as proteins and anatomical structures such as the cell membrane. Currently, ChEBI represents these biological activities of small molecules within a ‘role’ ontology which includes terms such as ‘cyclooxygenase inhibitor’. However, this ‘role’ ontology is not complete, and is not directly interlinked with the biological ontologies which serve as the main source of concepts describing biological entities. This makes it difficult to reason over the relationships between chemical entities and their biological targets. To address this issue, we propose a model for interrelating multiple ontologies and controlled vocabularies in the biomedical domain in order to formally characterise the bioactivity of small molecules. In support of this work, we have developed a method for analysing the scientific literature for textual descriptions of bioactivity events linked to chemical entities. We examine the distribution of terms from various controlled vocabularies (biological processes, proteins, organisms and organ/body parts) in combination with the chemical entities in the literature, to better understand reports of bioactivity. We find that proteins are the most commonly reported type of target of small molecule bioactivity, and that organisms and organs are most commonly reported in the literature as locational constraints rather than as targets.

1 Introduction

ChEBI is an ontology of chemical entities of biological interest [3]. It describes chemical entities such as molecules and ions together with their structural and biologically relevant properties. As of June 2011, it consists of around 25,000 entities, divided into a structure-based classification and a role-based classification. The role-based classification includes terms describing the biological activities of chemical entities, such as ‘cyclooxygenase inhibitor’ and ‘immunomodulator’. These terms describe small molecule *bioactivity*: the combined influence of a small molecular entity on the components of a living organism and on the organism as a whole.

On a molecular level, small molecule bioactivity corresponds to the binding of the molecule to a macromolecular receptor, resulting in some observable physiological effect on the biological systems involving that macromolecule [4]. Bioactive molecules can

have positive effects, such as repressing the development of disease, or they can have negative (toxic) effects, leading to illness or even death. The differentiation of bioactive molecules from non-bioactive molecules is one of the core requirements for *in silico* drug discovery approaches [12], as are delineating molecules which share similar activity profiles [9].

To properly formalise the description of activities of chemical entities in biological contexts requires reference to multiple terminological sources, some of which fulfill the requirements for formal ontologies (such as, e.g. the OBO Foundry ontologies [15]), whereas other ones are better characterised as thesauri, databases, or controlled vocabularies. For example, to formalise a description of enzymatic inhibitor activity requires reference to the enzyme which is being inhibited; to formalise participation in a particular biological process requires reference to the process; and bioactivity descriptions may require reference to the exact location of the

activity and the organism within which, or against which, the activity took place.

The ChEBI role ontology does allow the categorisation of chemical entities by their bioactivities. However, in its present form it suffers from two key limitations:

1. Role assertions are relatively *sparse* as compared to the full ontology of chemical entities (just less than 3000 chemical entities are mapped to just less than 500 roles, ca. 10% of the full chemical entity ontology). The result is that many of the chemical entities included in the ontology are not adequately described in terms of their biological context.
2. Bioactivity descriptions in the role hierarchy of the ontology are not explicitly linked to a primary reference source for the biological entities themselves. For example, the term ‘cyclooxygenase inhibitor’ describes the inhibition of a cyclooxygenase enzyme, yet this term is not explicitly linked to a reference for enzymes such as UniProt.

The aim of the present work is to use the automated analysis of literature as a means to address these limitations. The remainder of this paper is organised as follows. Firstly, we present our methods, which include the definition of a language model for bioactivity description and its application to extract mentions of bioactivity events from publicly available literature resources, in Section 2. Section 3 describes and discusses our results, including the implications of our literature analysis on the ontology model for interrelating chemical entities and biological objects and processes. In the final section we conclude with the relevance of this work both for biomedical research generally and for improved curation tools in the context of the ChEBI project.

2 Methods and Models

We first defined a language model for bioactivity terminology based on the examination of relevant portions of the Metathesaurus of the Unified Medical Language System (UMLS) [1] and the ChEBI biological roles. This is described further in Section 2.1.

We then used this language model to

extract bioactivity descriptions for ChEBI entities from MEDLINE abstracts. The text mining methods used are described in Section 2.2. After examining the sentences returned, we defined an ontology model for characterising the formal relationships between ChEBI entities and other biological entities.

2.1 Bioactivity Language Model

Basic Phraseal Patterns. Bioactivity of a chemical entity (CE) is described using given a set of language features: “inhibitor” and “activator”, “modulator”, “agonist” and “antagonist”, “toxin”, “regulator”, “suppressor”, “adaptor”, “stimulator”, “factor”, “messenger” and “blocker”; these will be called *trigger words*.

Any of these features can occur as a head noun in a phrase structure leading to the following type of phrasal patterns for their identification: a head noun preceded by a noun phrase, as follows: <modifier> <head>.

Ideally, the phrase composing (<modifier>) is constituted by one or more tokens which denote the *target* of the bioactivity, whereas the head word specifies the *nature* of the interaction between the small molecule and the target. For example, ‘beta-adrenergic receptor inhibitor’ has as modifier ‘beta-adrenergic receptor’ (the target) and as head word ‘inhibitor’ (the nature of the interaction is inhibition).

The basic language model was further extended to include alternative, compatible language patterns such as ‘inhibition of X’, where ‘X’ corresponds to the modifier and ‘inhibition’ the head word [8]. We identified four different syntactical structures for bioactivity descriptions, namely:

1. Noun phrase or adjective/adverb compositions as modifier. This is the most commonly seen structure of the basic noun phrase, and we find a considerable number of bioactivity terms presented in this way. For example:

HIV transcriptase inhibitor

2. Prepositional phrase as modifier. Prepositional phrases are generally formed by a preposition followed by a prepositional complement. We also find this structure is often represented in bioactivity terms. For example:

Suppressor of fused protein Oct-1
CoActivator in S phase protein, human

3. Verb phrase as noun phrase modifier. When the verb phrase functions as a modifier in a bioactivity noun phrase, it presents the way in which the activation of the described subject results in a kind of influence to its object. For example:

TIR domain containing adaptor
inducing interferon-beta protein

4. Relative clauses as modifier. Relative clauses are defined as subordinate clauses that consist of a clause beginning with a relative pronoun. This type of modifier is also used in the bioactivity presentation. For example:

Factor that binds to inducer of short
transcripts protein 1

2.2 Bioactivity Term Extraction from MEDLINE Abstracts

The method used to extract bioactivity descriptions from MEDLINE abstracts is a simple procedure which identifies noun phrase structures by matching *syntactical language patterns*. These hand-crafted language patterns form an alternative to syntactic parsing, which requires significant compute resources and is still error prone in several extraction tasks [7].

After bioactivity noun phrases were identified using the above patterns, we pruned outliers which had the trigger word as other parts of the phrase. For example, *Mononuclear cell growth inhibitor assay* is not considered to represent a valid bioactivity phrase because the activity term *inhibitor* is not the head noun (which in this case is *assay*). The solution for the identification of the noun phrases is based on hierarchically organised language patterns developed for the extraction of protein noun phrases in the protein-protein interaction pipeline [14]. The syntactical structures of the matching patterns have been tailored to fit the language model used in the approach of this manuscript.

The purpose of this analysis was to investigate the target types for bioactivity descriptions. To this end, four taggers for named entity and concept label identification (UniProtKB [10], Organ [6], Organisms [16] and GO [13]) were applied on the modifier of

candidates extracted from MEDLINE. The unique count of tagging was cross analysed by features provided in Section 2.1. We collected tabulated statistics which are presented in the results section.

Classifying Bioactivity Terms. After extracting bioactivity descriptions from MEDLINE, we aimed to find an efficient method of classifying all the candidates with a high rate of recall.

When the process results in the entire modifier being annotated by a tagger, this consequently indicates its semantic type. For example, *CaM kinase I activator* is easily classified as a *protein* activator since the modifier has been annotated as a protein from making reference to UniProtKB.

However, in the majority of cases, we found that the result is a nested case, in which the semantic tagger annotates just part of the modifier, i.e. the tagged result resides within the boundaries of the whole phrase for the modifier. For instance, *Agkistrodon blomhoffi ussuriensis protein C activator*. In this case, *ussuriensis protein C* is the authentic target of the activation, though *Agkistrodon blomhoffi* is identified. As previously mentioned, a simple method to rule out un-associated tagging is used. We retain the tag which is in the last position within the modifier, ignoring other tags. In this example, the target type is not species but protein.

2.3 Text Mining Methods for Bioactivity Triple Extraction

We used a dictionary-based approach to extract names of small molecules and their targets together with their relation types from the whole MEDLINE resource. The approach processed text on a sentence level, extracting triples which contain (1) a small molecule term, (2) the 'feature' trigger word, which presents the relation type, and (3) a term representing the target of the small molecule.

To identify the small molecules, we compared results from using two different chemical taggers, namely the newest versions of Oscar3 [2] and Jochem [5]. Jochem, being dictionary-based, has the advantage that all chemical entities it recognises are known entities, whereas Oscar3 can recognise non-known strings that resemble syntactical

structures denoting chemical entities (higher recall).

All the possible combinations of small molecule terms, features and target terms in each sentence are generated. We found that false positive cases were significant, and therefore applied three stages of rule-based filtering:

1. Remove triples from the candidate list when the putative small molecule term is actually a role term according to the ChEBI ontology (e.g. 'antibiotic')
2. Filter out those triples where the small molecule term has the suffix "-ase", since these terms are normally enzyme names.
3. Remove triples when the string that supposedly denotes a small molecule has less than three characters.

3 Results and Discussion

3.1 Evaluation of Language Model

The evaluation of our approach is ongoing work and requires a gold standard corpus (GSC). The GSC would enable us to test supervised learning methods against our existing feature-based extraction method. However, the named entity recognition methods have all been evaluated. The identification of proteins and genes performs at 52.37%/61.63%/56.62% (Rec,Prec,F-Meas) on PennBioIE and 50.26%/61.63%/56.62% (Rec,Prec,F-Meas) on BC-II [11]. The method applied for the identification of genes and proteins was based on the UniProtKB dictionary with basic disambiguation and was not trained on one of the different gold standard corpora, since methods trained on gold standards show high differences in their performance when being tested against other gold standard corpora. Trained methods for gene mention identification are available and show higher performance, but do not allow linking results to data from biomedical data resources, e.g. UniProtKB and EntrezGene.

3.2 Results of Running Language Model on MEDLINE Abstracts

Table 1 shows an overview of target type associated with feature trigger words. Each

cell shows the unique count of semantic tagging for a certain feature. Both nested and exact matching on the modifier of bioactivity terms are considered.

Feature	Protein	Organ	Organism	Biological Process
stimulator	2,526	3,303	500	1,808
adaptor	3,729	100	133	1,016
modulator	7,847	1,468	536	4,204
messenger	10,056	1,186	1,151	3,876
agent	10,522	10,292	19,374	8,744
blocker	13,588	1,371	9,235	4,203
toxin	16,890	1,583	10,265	3,276
suppressor	18,534	1,301	2,382	2,988
regulator	27,724	5,469	2,802	27,270
factor	40,427	21,959	11,152	77,670
agonist	48,973	3,633	13,154	12,353
activator	71,165	1,745	3,895	19,376
antagonist	80,932	9,483	11,740	19,486
inhibitor	336,420	12,102	30,839	142,289

Table 1. Identifying target type of small molecule on MEDLINE abstracts.

From this table, the main target types of bioactivity are identified based on a two-feature driven method.

In general, protein names are mostly nested in the modifier of bioactivity terms. UniProtKB tagging and 'inhibitor' gives a high number of hits: 336,420 unique combinations. This suggests that bioactivity descriptions in text usually refer to activities against a protein or enzyme. Two such examples are:

- *Other lysosomal hydrolases are not inhibited by N-bromoacetyl-beta-D-galactosylamine, with the exception of 'neutral' beta-glucosidase glucosylhydrolase.*
- *At the biochemical level cardiac guanylate cyclase activity is enhanced 2-3 times with acetylcholine and this enhancement is completely blocked by atropine.*

There are not as many hits in the Organ and Organisms groups. We can find a few true positive examples such as *bothrops jararaca* inhibitor and *thyroid stimulator*. However, there are many examples in which the organ or organism appears in the sentence only to denote the location of the bioactivity being described. For example:

1. Caesium ion antagonism to chlorpromazine - and L-dopa-produced

behavioural depression *in mice*.

2. The changes in the contents of glycolytic intermediates in the livers indicate that the phosphoenolpyruvate carboxykinase [EC 4.1.1.32] reaction is inhibited by tryptophan administration *in all groups of rats*.
3. The oral administration of meta-proteranol increased the leukocyte adenylyl cyclase activity which was stimulated by NaF and decreased the count of peripheral eosinophils *in some of the monkeys*.

We conclude that in the literature, organ and organism most commonly provide the contextual information about where a bioactivity takes place, rather than being themselves the target of the bioactivity. This will influence our ontology model, described in Section 3.4.

We also analysed the case where GO terms were tagged in bioactivity terms. For example, *inhibitor of DNA transcription*. Here, a biological process is the target of the bioactivity term.

Limitations. As is the norm in this type of text mining approach, there are also typical ‘noisy’ false positives in the result, such as ‘hand’ being tagged as a body part in the sentence ‘On the other hand, ...’, and ‘dialysis’ being tagged as a species in the sentence ‘Influence of peritoneal dialysis on factors affecting oxygen transport.’ (Dialysis is, indeed, a species: a kind of bug.). Care also needs to be taken in that some of the results reflect sentences in which the bioactivity being described in the extracted triple is explicitly *not* reported as taking place, such as:

1. *Without* influence on WDS were: physotigmine, atropine, ganglionic-or adrenergic-blocking drugs, Dopa, MAO-inhibitors, serotonin- and histamin-antagonists and nonnarcotic analgesics.
2. The cellulase component was *not* markedly inhibited by most metal ions tested.

3.3 Comparison of Chemical Taggers

To identify chemical entities, we compared a dictionary-based approach using Jochem with the results generated using Oscar3 which is able to identify novel chemical names in text using a machine learning approach.

Table 2 shows the frequency of each triple mentioned in text together with the unique count of triples before and after the rule-based filtering described in Section 2.3.

Oscar3 yields many more triples than Jochem does. This is expected, since Oscar3 recognises any chemical-like string. However, Oscar3’s approach also results in a considerable number of false positives due to its recognition of chemical-like nomenclature appearing as a component in larger strings (such as protein names). Furthermore, we can observe a smaller number of triples identified by UniProtKB and Oscar3 compared to the set identified by UniProtKB and Jochem. This is because Oscar3 produces annotations that nest within a protein mention in the sentence and thus lowers the subsequent annotation protein mentions. Jochem performs more long-form matching than Oscar3 does, therefore the following protein identification has a higher likelihood of identifying a protein term within the sentence, hence yielding a greater number of triples.

The comparison of before and after filtering show whether the triple mention is by chance and the association between the chemical and the other semantic group is more than contextually related. Between chemicals and proteins the ratio is smaller than the other groups. The non-unique number of triples is less than twice the number of unique ones, while it is more than this ratio in other groups (specifically in the chemical organ group). The number of non-unique triples identified by Jochem after filtering is almost three times the unique count.

		UniProtKB		Organ		Organism		GO	
Chemical tagger	Filtering	uniq	non uniq	uniq	non uniq	uniq	non uniq	uniq	non uniq
Jochem	before	4,114,286	7,853,314	2,666,468	7,148,677	1,785,771	4,076,253	1,244,099	2,947,289
	after	2,912,756	5,457,529	1,632,855	5,302,115	1,394,310	3,085,056	935,864	2,089,163
Oscar3	before	11,599,131	23,988,686	4,344,247	11,855,944	2,672,206	5,836,725	1,864,403	4,607,315
	after	7,827,737	12,776,542	2,222,450	4,598,353	1,347,442	2,338,487	945,320	1,804,411

Table 2. Triples analysis from MEDLINE

3.4 Ontology Model for Interrelating Small Molecules and Biological Entities

The relationship between the chemical entity and its bioactivity which is already used in ChEBI is *has_role*.

Based on our analysis of bioactivity phrases in the literature, we have identified macromolecules and biological processes as the most common types of targets for the bioactivity of small molecules. We could therefore introduce a *has_target* relationship to relate a bioactivity description to either a macromolecule or a biological process. However, strictly speaking, the range of the *has_target* relationship should be restricted to those entities with which the chemical entity can physically interact – macromolecules. We can assume that biological processes are mentioned where the exact macromolecular target is unknown. In the same way, anatomical or subcellular locations may be mentioned when the exact target is unknown. Therefore, we can further formalise the *has_target* relation link to processes: in this case the target is a *macromolecule and participant_of some Process* (Manchester syntax).

Examples:

```
m1 is a betaadrenergic receptor inhibitor:
m1 subclassOf bearer_of some
  (realized_by only
    (Inhibition and
      (has_target some BetaAdrenergicReceptor)))

m2 is a mitosis stimulator:
m2 subclassOf bearer_of some
  (realized_by only
    (Stimulation and
      (has_target some
        (participant_of some Mitosis))))

m3 is a thyroid stimulator:
m3 subclassOf bearer_of some
  (realized_by only
    (Stimulation and
      (has_target some
        (has_locus some ThyroidGland))))

m4 is a mouse thyroid stimulator:
m4 subclassOf bearer_of some
  (realized_by only
    (Stimulation and
      (has_target some
        (has_locus some (ThyroidGland and
          part_of some Mouse))))))
```

We have noted that organisms, organs and bodily parts appear frequently as contextual, locational modifiers for the bioactivity descriptions in the literature. In these cases, the above formalism is too strict, since the location is assumed to contribute to the definition of the bioactivity. We therefore introduce an additional relationship, *has_context*, which may hold between a bioactivity description and an organism, bodily organ or component to express *non-definitional* information: the bioactivity *can* take place in many organisms, but was *discovered* through investigations in one specific organism.

An important limitation of Description Logic-based ontology representation formalisms is that they are unable to elegantly express the fact that the context applies not to a bioactivity description *per se*, but rather to a small molecule-bioactivity association. This would require a ternary relationship. However, for our purposes it will be sufficient to assume that we can get around this problem through the standard method of reification.

Finally, we note that the different head nouns used in our analysis (inhibitor, antagonist and so on) correspond to different types of bioactivity, such as are delineated by upper-level distinctions in the ChEBI role ontology.

4 Conclusions

We have presented a language model for bioactivity descriptions which we have used to examine the distribution of bioactivity descriptions in the scientific literature. From this analysis we derive insights into the model needed to accurately formalise an ontology for bioactivity, appropriately distinguishing between bioactivity targets and contextual (locational) information. Such an ontology will serve as a bridge between small molecules, their biological targets, and the locations and contexts in which they act, allowing automated reasoning about the activities of chemical entities in a biological context.

This work should be understood as a first step in the direction of such a formalisation, a pressing goal in the context of ChEBI's participation in the OBO Foundry effort to interrelate ontologies in the biomedical domain. Future work will develop our text analysis platform further as a support utility

for ChEBI curation, and aim to incorporate the increased formalisation described here directly into the ChEBI ontology. Since ChEBI is a manually curated resource, we cannot pre-populate ChEBI with extracted relationships based on the text mining methods described here. However, such automatically identified bioactivity descriptions in the literature can be used to provide semantically enriched information in our ontology curation workbench, which allows a much improved and more rapid curation experience.

References

1. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucl. Acids Res.* 32 (suppl 1), D267–270 (Jan 2004).
2. Corbett, P., Murray-Rust, P.: High-Throughput Identification of Chemistry in Life Science Texts, *Lecture Notes in Computer Science*, vol. 4216, chap. 11, pp. 107–118. Springer, Berlin (2006).
3. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. *Nucl Acids Res* 36(suppl 1), D344–D350 (2008).
4. Gohlke, H., Klebe, G.: Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed* 41, 2644–2676 (2002)
5. Hettne, K.M., Stierum, R.H., Schuemie, M.J., Hendriksen, P.J.M., Schijvenaars, B.J.A., Mulligen, E.M., Kleinjans, J., Kors, J.A.: A dictionary to identify small molecules and drugs in free text. *Bioinformatics* 25(22), 2983–2991 (Nov 2009).
6. Hishiki, T., Ogasawara, O., Tsuruoka, Y., Okubo, K.: Indexing anatomical concepts to omim clinical synopsis using umls metathesaurus. In *Silico Biology* 4 (2003)
7. Hobbs, J.R., Appelt, D.E., Bear, J., Israel, D.J., Kameyama, M., Stickel, M.E., Tyson, M.: Fastus: A cascaded finite-state transducer for extracting information from natural-language text. *CoRR* [cmp-lg/9705013](https://arxiv.org/abs/1907.05013) (1997)
8. Kirsch, H., Gaudan, S., Rebholz-Schuhmann, D.: Distributed modules for text annotation and IE applied to the biomedical domain. *International Journal of Medical Informatics* In Press.
9. Lipinski, C., Hopkins, A.: Navigating chemical space for biology and medicine. *Nature* 432 (2004)
10. Magrane, M., Consortium, U.: UniProt knowledgebase: a hub of integrated protein data. *Database: the journal of biological databases and curation* 2011 (Mar 2011).
11. Morgan, A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., et al.: Overview of biocreative ii gene normalization. *Genome biology* 9 (Suppl 2), S3 (2008)
12. Oprea, T.I., Tropsha, A.: Target, chemical and bioactivity databases – integration is key. *Drug Discovery Today: Technologies* 3, 357–365 (2006)
13. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Yepes, A.J.: Text processing through Web services: calling Whatizit. *Bioinformatics (Oxford, England)* 24(2), 296–298 (Jan 2008).
14. Rebholz-Schuhmann, D., Jimeno-Yepes, A., Arregui, M., Kirsch, H.: Measuring prediction capacity of individual verbs for the identification of protein interactions. *Journal of biomedical informatics* (Oct 2009).
15. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25(11), 1251–1255 (Nov 2007).
16. Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., Rapp, B.A.: Database resources of the NCBI. *Nucl acids res* 28(1), 10–14 (Jan 2000).

A Meta-Data Approach to Querying Multiple Biomedical Ontologies

Ravi Palla^{1,2}, Dan Tecuci¹, Vinay Shet¹, Mathaeus Dejori¹

¹Siemens Corporate Research, Princeton, NJ, USA

²School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

Abstract. We present an approach for retrieving information spread across multiple large ontologies, with the goal of developing a biomedical question answering system that can assist physicians in diagnosis, treatment and therapy planning. The approach involves ontology integration and run-time SPARQL query generation, both of which are accomplished by defining a meta-ontology containing information about the properties and structure of the individual ontologies. The approach enables ontology integration with minimal changes and also supports ontology interoperability. We built a prototype of our approach that integrates the Foundational Model of Anatomy ontology, the human disease ontology, and an ontology that represents certain information from the Merck manual.

1 Introduction

Biomedical ontologies are rich sources of information that can be shared and used for reasoning within question answering (QA) systems. To assist physicians in making the correct diagnosis and prescribing the right medication, a QA system needs to have sufficient access to information regarding anatomy, pathology, pharmacology, and other related domains. While there are ontologies that cater to each of these individual domains, to the best of our knowledge, there is no ontology that sufficiently covers all these domains. Moreover, the ontologies for the individual domains do not necessarily contain all the information required by QA systems to effectively assist physicians. Therefore, such systems need to both enhance and integrate several biomedical ontologies.

In this paper, we present an approach for retrieving information spread across multiple ontologies in the context of building a question answering system. It involves ontology integration and run-time SPARQL query generation, both of which are accomplished by defining a meta-ontology that contains information about the various properties in the ontologies, the mapping between the properties, and the information needed to generate SPARQL queries for retrieving information with respect to these properties.

The approach abstracts away the actual ontologies as it refers only to the meta-ontology to retrieve the required information. This implies that ontologies can be integrated into the system by simply updating the meta-ontology. The approach also allows for interoperability between ontologies at the level of ontology alignment [1]. While this is a weak form of integration, we found it to be appropriate for QA systems that rely on several large ontologies. We tested our approach by considering the Foundational Model of Anatomy (FMA) ontology¹ [3], the human disease ontology², and an ontology that represents certain information from the Merck manual³. The metadata we use is tailored for QA systems, but the approach itself can be used for other applications.

2 The Approach

We define a set of high-level properties based on the types of questions to be answered and create an upper ontology that maps these high-level properties to properties of individual ontologies.⁴ The mapping between the properties of individual ontologies can be derived from

¹ <http://sig.biostr.washington.edu/projects/fm/>

² <http://www.obofoundry.org/>

³ <http://www.merckmanuals.com/professional/index.html>

⁴ This mapping is not necessarily one-to-one and can be a many-to-many.

their mapping to the high-level properties. Consider the case of answering questions of the form, “What is a [concept-name]?” For this, we define a high-level property “definition” and then update the upper ontology by including the mapping between “definition” and the properties of the individual ontologies that provide an appropriate answer for a definitional question.

Given this, in order to retrieve the URIs of all properties necessary to answer definitional questions, the system can simply query the upper ontology. However, just retrieving these URIs is not sufficient to answer the question and correct queries need to be formulated to retrieve the definitions. In order to formulate the queries, the knowledge of the structure of the underlying ontologies is required, information that can also be included in the upper ontology. The following RDF description shows a possible mapping from the high-level property “definition” to the property providing definitions in the disease ontology. The description also contains the necessary information to generate the SPARQL queries that can be used to retrieve the definitions.⁵

```
<rdf:Description
rdf:about="OBOINOWL#hasDefinition">
  <hasProperty>definition</hasProperty>
  <hasQueryTarget>def</hasQueryTarget>
  <hasQueryLine>?x. ?x rdfs:label
    ?def.</hasQueryLine>
</rdf:Description>
```

The description above indicates that the SPARQL queries needed to retrieve definitions from the disease ontology are of the form

```
SELECT ?def WHERE {
[subject] <OBOINOWL#hasDefinition> ?x.
?x rdfs:label ?def.
}
```

where “[subject]” is the URI of the concept in the disease ontology whose definition has to be retrieved.

2.1 Interoperability: Handling Synonyms

The approach presented so far assumes that all the URIs corresponding to the concept names in the user’s question have been identified. However, this is not trivial since

different ontologies might refer to a concept using different names, and the user can use any of these names in the question. For example, the Merck manual ontology contains the definition for “Atrioventricular block” and the user can ask the question, “What is a AV block?” The disease ontology contains the name “AV block” as a synonym for “Atrioventricular block” but does not contain the definition. So, if the system only gets the URIs corresponding to the name “AV block”, it will not be able to answer the question. In order to answer the question, the system needs to retrieve the URIs corresponding to “Atrioventricular block” in the Merck manual ontology, and this can be done by first retrieving the synonyms of “AV block” from the disease ontology and then using them to obtain the corresponding URIs from the Merck manual ontology. In general, to answer any question about a concept, the system needs to first retrieve all the synonyms of the concept name used in the question and then use them to retrieve the corresponding URIs.

However, since different ontologies have different structures, querying for the synonyms is not straightforward. To address this, we define a high-level property “synonym” and use the upper ontology to represent information about querying for synonyms. The following description shows one way to represent information about retrieving synonyms from the disease ontology.

```
<rdf:Description
rdf:about="OBOINOWL#hasExactSynonym">
  <hasProperty>synonym</hasProperty>
  <hasQueryTarget>syn</hasQueryTarget>
  <hasQueryLine>?x. ?x rdfs:label
    ?syn.</hasQueryLine>
</rdf:Description>
```

In order to obtain the synonyms, the system can query the upper ontology to retrieve all the information required to formulate the SPARQL queries needed to retrieve the synonyms.

3 The QA System, High-Level Properties and their Mappings

We have implemented a prototype QA system that uses this approach to query multiple biomedical ontologies. The system answers questions of the form “What is a [concept-

⁵ Here, OBOINOWL is an abbreviation for “<http://www.geneontology.org/formats/oboInOwl>”.

name]?” and “What is/are the [relation-name(s)] of the [concept-name]?” With respect to such questions we defined high-level properties like “definition”, “part”, “location”, “connections”, and “affected organs”, so that questions such as “What is the location of the heart?” and “What are the affected organs of atrial fibrillation?” can be asked.⁶

The table below shows some of the mappings for the high-level properties discussed above.

High-Level Property	Mapped To
definition	fma:definition, obolnOwl:hasDefinition, fma:location, merck:hasDefinition, fma:surrounded_by, rdfs:subClassOf
location	fma:location, fma:surrounded_by, fma:contained_in

The table suggests that in order to answer definitional questions of the form “What is a [organ]?”, the system also retrieves information about the type of the organ and some information about the location of the organ. This is another advantage of our approach as it enables us to change the information that is retrieved by simply adding/deleting certain mappings.

4 Related Work and Conclusion

Building a comprehensive biomedical QA system requires some level of integration of multiple domain ontologies. There have been several approaches presented for integrating biomedical ontologies, like the Ontology of Biomedical Reality (OBR) framework [2] and the framework of the Open Biomedical Ontologies (OBO)⁷ consortium that attempt to make the process of development of biomedical ontologies more formal, thereby allowing more interoperability between ontologies spanning several domains. Among other approaches, is

the Linked Life Data⁸ platform that has been used to integrate biomedical ontologies spanning multiple domains. While our approach also deals with ontology integration, albeit weak, it differs from the above approaches in that it is goal driven. We define a set of high-level properties of interest based on the kinds of questions we want the system to be capable of answering and do the integration with respect to these properties.

While our approach is still in its infancy, preliminary results show that it enables ontologies to be integrated into the QA system with minimal changes while supporting interoperability between ontologies with different structures as shown in section 2.1. While the interoperability supported is a weak form of integration, it seems appropriate for QA systems that need to consider several large ontologies of different structures.

We plan to extend the QA system to handle more question types, answering which would not only require retrieval of information but also deep reasoning so that the system can effectively assist physicians in their tasks.

Acknowledgements

We are grateful to Michael Sintek and Patrick Ernst from the German Research Center for Artificial Intelligence (DFKI) for their contribution to this work, and to the anonymous referees for their valuable comments.

References

1. Sowa, J.F.: Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks/Cole (2000).
2. Rosse, C., Kumar, A., Mejino, J.L.V. Jr., Cook, D.L., Detwiler, L.T., and Smith, B.: A Strategy for Improving and Integrating Biomedical Ontologies. In Proceedings of AMIA Symp. (2005) 639–643.
3. Rosse, C., Mejino, J.L.V. Jr.: A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J. Biomed. Inform. 36(6) (2003) 478–500.

⁶ The information regarding the affected organs is retrieved from an ontology obtained by a stronger integration of the FMA and the disease ontologies.

⁷ <http://www.obofoundry.org/>

⁸ <http://linkedlifedata.com/>

Multiple Ontologies in Healthcare Information Technology: Motivations and Recommendation for Ontology Mapping and Alignment

Colin Puri¹, Karthik Gomadam¹, Prateek Jain², Peter Z. Yeh¹, Kunal Verma¹

¹Accenture Technology Labs, San Jose, CA, USA

²Kno.e.sis Center, Wright State University, Dayton, OH, USA

Abstract. Electronic Health Records (EHR), Personal Health Records (PHR), data analysis and integration have emerged as key pieces in the delivery of quality health care. Integration of heterogeneous sources of patient information, domains of healthcare information, and associated ontologies brings about important questions. This paper enumerates upon some of the issues of ontology alignment, mapping, and motivations for the need of integration with respect to patient health care. No single ontology is sufficient to meet the growing needs of today's healthcare and the ontologies that exist today must themselves be integrated together for support of data integration and analysis. We also make a recommendation on one potential solution.

Keywords: Biomedical Ontologies, Ontology Mapping, Ontology Alignment, Healthcare Information Technology, BLOOMS

1 Introduction

As the healthcare industry moves towards wider adoption of Healthcare Information Technology (HIT) solutions such as Electronic Health Records (EHRs) and Personal Health Records (PHRs), data analysis and integration has emerged as a significant component in the delivery of quality healthcare services. For example, patient data can come in EMR systems, which capture treatments, symptoms, and diseases. It can also come from PHRs, such as Google Health¹ and Microsoft HealthVault², and other health and wellness applications, such as LiveStrong³ and TrialX⁴, that capture additional aspects of the patient's health such as lifestyle choices and diet. If these data sources are properly integrated, then we can begin to realize applications that will enable healthcare providers to effectively answer questions such as:

1. What treatments were administered to other patients with similar health conditions?

2. What was the efficacy of such treatments when administered to patients with a given physiological profile?
3. What medications are currently being prescribed to the patient and how do they constrain available treatment options?
4. How can one meaningfully find and utilize the vast amounts of medical knowledge, such as codified medical vocabularies, scientific publications, and findings from clinical trials, available in the public domain?
5. How can the health and wellness information stored by a patient in PHRs and other PHR-based applications be used to improve the quality of care?

Such applications can potentially save billions of dollars in healthcare costs [12], while improving the quality of care.

Biomedical ontologies provide a promising solution for integrating these heterogeneous data sources by providing a common vocabulary (and framework) to enable interoperability, resolve ambiguity, etc. However, no single ontology is sufficient. Instead, multiple ontologies must be combined in practice to fully

¹ <http://health.google.com>

² <http://healthvault.microsoft.com>

³ <http://livestrong.com>

⁴ <http://trialx.com>

realize meaningful integration and analysis of data in the healthcare domain. Hence, the ontologies themselves must first be integrated before they can support the necessary data integration and analysis.

This paper focuses on the technical challenges in integrating multiple ontologies, and takes the position that existing ontology alignment solutions can provide a viable solution to this end.

2 Ontology Mapping and Alignment: Motivation and Current Approaches

A patient's medical record captures multiple aspects of his/her health (e.g. medications, health conditions, prior treatments, etc) and can come from multiple sources (e.g. EMR systems, PHR applications, etc). Integrating this information into a coherent view requires combining multiple ontologies such as:

- SNOMED CT [16] is an systematic organization of medical terminology containing information related to medical conditions, procedures, pharmaceuticals, etc.
- RxNorm [7] provides a vocabulary for normalized names for clinical drugs. It is intended to cover all prescription medication in the United States. It contains the active ingredients, strengths, and dose form comprising that drug.
- MeSH (Medical Subject Headings) [3] is a large and expansive controlled vocabulary for indexing medical journals, articles, and books.
- ICD-10 (International Statistical Classification of Diseases and Related Health Problems 10th Revision) [2] is a collection of codes specifying diseases, symptoms, findings, complaints, etc. as defined by the World Health Organization (WHO).
- Gene Ontology (GO) [4] is a unified ontology designed to represent the gene attributes across all species. Furthermore, the goal of this ontology effort is to develop a controlled vocabulary, annotate gene information, and provide a set of useful tools for access to the

genetic information. The GO is a part of a larger initiative, the Open Biomedical Ontologies, to create controlled vocabularies for use between several biomedical domains.

A number of efforts such as UMLS [6], OpenGALEN [14], and 3M's Health-Care Data Dictionary [1] have tried to consolidate multiple ontologies. While these efforts do provide mappings between different biomedical ontologies, health-care providers still face several challenges when integrating their proprietary vocabulary and processes with third-party biomedical ontologies. These challenges range from syntactic differences (e.g. different terminologies, naming conventions, and formats) to deeper semantic differences (e.g. different granularity for modeling steps in a medical protocol). What is required are solutions that can generate these mappings either automatically or with minimal human effort.

Ontology mapping and alignment has been an active area of research. Various strategies, including machine learning, rule based mapping, and logic driven frameworks, have been adopted to address the challenge of ontology mapping. We briefly illustrate some of the research that have employed these techniques. Machine learning approaches have been used in Learning Source Description (LSD) [10]. LSD employs a multi-stage learning approach and exploits both the schema and the data. The Ontology Integration System (OIS) [8] adopts a query based approach and employs description logic based techniques. A hybrid approach, employing rules and learning is discussed in [11]. In addition to these techniques, ideas from the area of database schema matching have also been adopted in the context of ontology mapping. A survey of such approaches is presented in [15].

In general, ontology mapping can be classified into three categories [9]:

1. Global ontology view to local ontology view: An example of this would be the mappings between an ontology describing a provider's proprietary terminology and clinical pathways that use a view of SnoMed, with SnoMed.
2. Semantic mappings between local and target entities: An example would include

the mappings between an ontology for drug formulary and an ontology for clinical pathways, where a drug (a source entity) is mapped onto a medication (target entity). Once the mapping is done, the transformed entity captures the properties of the drug from the source ontology and the dosage information for a particular medical condition from the target ontology.

3. Mappings to enable ontology re-use by integration and alignment: An example would be the mappings to integrate multiple clinical pathway ontologies of similar chronic medical conditions that will help in identifying overlapping concepts and synonyms.

3 Recommendation

Approaches (described in the previous section) have shown promising results, but their application has been limited to a few public ontologies. There is one approach – i.e. BLOOMS [13] – which has been successfully applied to aligning disparate ontologies in the Linked Open Data Cloud [5] – a Web-scale effort to integrate vocabulary from diverse sources and providers, ranging from music ontology to consumer business categories. We recommend the consideration of BLOOMS as a possible engine for aligning disparate biomedical ontologies.

BLOOMS is a system for generating links between class hierarchies between two ontology schemas. In the context of the Linked Open Data Cloud (LODC), BLOOMS uses Wikipedia to create an initial set of categories for the given concepts and uses the comparison of the generated categories as a basis for link generation. Unlike many of the current approaches, BLOOMS uses the data resources available on the Web as a point of reference during the tasks of mapping and alignment. For example, BLOOMS uses Wikipedia to derive a category hierarchy for the concepts in the ontologies. The ability of the system to identify and leverage on non-traditional and open data sources makes it a flexible framework for alignment, while also reducing the dependency on the domain models. The latter advantage is significant and allows BLOOMS to deliver higher quality mappings.

References

1. 3M Healthcare Data Dictionary (3M HDD). <http://www.3mtcs.com/products/hdd>.
2. International Classification of Diseases, 10th Revision. <http://www.who.int/classifications/icd/en/>.
3. MeSH (Medical Subject Headings). <http://www.ncbi.nlm.nih.gov/mesh>.
4. The Gene Ontology. <http://www.geneontology.org/>.
5. C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
6. O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267, 2004.
7. O. Bodenreider. Visualization tools for the Unified Medical Language System (Sem-Nav), the Gene Ontology (GenNav), and RxNorm. In *Humans and the Semantic Web HCIL Workshop*, 2006.
8. D. Calvanese, G. De Giacomo, and M. Lenzerini. A framework for ontology integration. In *The Emerging Semantic Web Selected Papers from the First Semantic Web Working Symposium*, pages 201–214, 2002.
9. N. Choi, I. Song, and H. Han. A survey on ontology mapping. *ACM Sigmod Record*, 35(3):34–41, 2006.
10. A. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: A multistrategy approach. *Machine Learning*, 50(3):279–301, 2003.
11. M. Ehrig, S. Staab, and Y. Sure. Bootstrapping ontology alignment methods with APFEL. *The Semantic Web–ISWC 2005*, pages 186–200, 2005.
12. R. Hillestad, J. Bigelow, A. Bower, F. Girosi, R. Meili, R. Scoville, and R. Taylor. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs*, 24(5):1103, 2005.
13. P. Jain, P. Hitzler, A. Sheth, K. Verma, and P. Yeh. Ontology alignment for linked open data. *The Semantic Web–ISWC 2010*, pages 402–417, 2010.
14. A. Rector, J. Rogers, P. Zanstra, and E. Van Der Haring. OpenGALEN: open source medical terminology and tools. American Medical Informatics Association, 2003.
15. P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics IV*, pages 146–171, 2005.
16. K. Spackman, K. Campbell, et al. SNOMED CT: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association, 1997.

Modularization for the Cell Ontology

Christopher J. Mungall¹, Melissa Haendel², Amelia Ireland¹, Shahid Manzoor¹,
Terry Meehan³, David Osumi-Sutherland⁴, Carlo Torniai², Alexander D. Diehl⁵

¹Lawrence Berkeley National Laboratory, Berkeley, CA, USA

²Oregon Health and Sciences University, Portland, OR, USA

³The Jackson Laboratory, Bar Harbor, ME, USA

⁴The University of Cambridge, Cambridge, UK

⁵University at Buffalo, Buffalo, NY, USA

Abstract. One of the premises of the OBO Foundry is that development of an orthogonal set of ontologies will increase domain expert contributions and logical interoperability, and decrease maintenance workload. For these reasons, the Cell Ontology (CL) is being re-engineered. This process requires the extraction of sub-modules from existing OBO ontologies, which presents a number of practical engineering challenges. These extracted modules may be intended to cover a narrow or a broad set of species. In addition, applications and resources that make use of the Cell Ontology have particular modularization requirements, such as the ability to extract custom subsets or unions of the Cell Ontology with other OBO ontologies. These extracted modules may be intended to cover a narrow or a broad set of species, which presents unique complications. We discuss some of these requirements, and present our progress towards a customizable simple-to-use modularization tool that leverages existing OWL-based tools and opens up their use for the CL and other ontologies

1 Introduction

Many bio-ontologies were initially conceived of as stand-alone monolithic entities, developed independently of other ontologies. However, a modular approach, whereby portions of other ontologies are reused and made interoperable has many advantages [23], and this was one of the reasons for the establishment of the OBO Foundry [14, 1, 25]. With a modular approach, more complex ontology classes are constructed combinatorially using simpler ontology classes as building blocks. These building-block classes may come from separate ontologies, or from orthogonal hierarchies within a single ontology. For example, a cell type such as *mature eosinophil* in the OBO Cell Ontology (CL) [2] can be functionally defined using the biological process class *respiratory burst* from the Gene Ontology (GO) [20]. Inter-ontology dependencies such as these can be bi-directional; for example, a GO process such as *eosinophil differentiation* can be defined in terms of the CL class *eosinophil* [21]. The CL is also taxonomically modularized in that it leaves representation of highly specialized cell types to species-specific ontologies. (In this manuscript we use the

term *taxon* in the sense of an organism/species taxonomy.)

Table 1 shows the external ontologies that are of relevance to the CL modularization strategy.

Module Extraction

When working in the context of multiple ontologies, it is important to be able to extract sub-modules from combinations of ontologies. For example, when working with the CL, it can be useful to extract the minimal subset of GO that is required to perform automated reasoning over the CL and obtain results that are valid and complete. This subset can either be *imported* or *merged* into the source ontology. If the entire external ontology is imported or merged without first extracting a subset, the resulting ontology union can be difficult to work with and reason over.

Module extraction is also useful for downstream applications, such as using an ontology in annotation or analysis. Annotators may want a subset of the ontology that is of relevance to their taxon or domain of interest, and term enrichment tools benefit from using a subset as it decreases the size of the hypothesis space, resulting in improved p-values. Since its

inception, the GO has catered to these use cases by providing manually created subsets or “GO slim” [12]. Using “slimming tools” (Ireland, unpublished), GO annotations can be mapped from a full ontology to a slim.

Ontology	Scope and Relevance to CL
PR [22]	<i>The Protein Ontology</i> Proteins from multiple species - Used to define cell types based on presence of specific proteins
GO [12]	<i>The Gene Ontology</i> Biological Processes, Molecular Functions, and Cellular Components - Biological Processes used for defining cells by function, and Cellular Component for surface receptors
PATO [9]	<i>The Ontology of Phenotypic qualities</i> Qualities that can apply to anatomical entities or biological processes used to define cells in terms of shape and other physical characteristics.
UBERON [11]	<i>The Uber anatomy ontology</i> Gross anatomical structures spanning metazoa (but like CL, with a significant vertebrate bias) - used to define cells by location in the organism
NCBI Taxon	<i>Taxonomy of species</i> For taxonomic constraints. Only a very small subset is required.
MA [13]	<i>The adult Mouse Anatomy ontology</i> Species-specific gross anatomy
FBbt [8]	<i>The Drosophila anatomy ontology</i> Species-specific gross and cellular anatomy
WBbt [19]	<i>The C. elegans anatomy ontology</i> Species-specific gross, cellular and subcellular anatomy
FAO [3]	<i>The Fungal Anatomy Ontology</i> Multi-species gross, cellular, and subcellular anatomy
ZFA	<i>The Zebrafish Anatomy ontology</i> Species-specific gross and cellular anatomy
TAO [5]	<i>The Teleost Anatomy ontology</i> Multi-species gross and cellular anatomy
PO [16]	<i>The Plant Structure ontology</i> Multi-species gross and cellular anatomy
FMA [24]	<i>The Foundational Model of Anatomy - adult human</i> Species-specific gross, cellular, and subcellular anatomy

Table 1. Ontologies required by the Cell Ontology for importing modules and/or coordinating development.

Manually creating subsets is a time-consuming task, and will not scale for all

purposes, so automated techniques are extremely valuable. The problem of extracting minimal subsets that preserve reasoner results has received considerable attention in the Description Logic literature – see for example [17]. The majority of the discussion has been on the theory; some module extraction tools have been implemented for OWL ontologies, but they are not always easy to use.

The MIREOT (Minimal Information for Retrieval of an External Ontology Term) guidelines [4] and associated tooling [27] provide support for practical module extraction. One notable feature of MIREOT is that external ontology axioms are typically merged into the source ontology rather than imported, potentially leading to synchronization issues. MIREOT has been adopted by some ontologies, such as the Ontology of Biomedical Investigations (OBI) and eagle-i [10]. The CL is currently using a MIREOT strategy with the ontology editor OBO-Edit [6] to create an extended version of the CL which includes externally referenced classes. These are removed for the “basic” version of the CL.

Taxonomic Module Extraction

Another requirement is to extract sub-modules from unions of ontologies. For example, cross-species comparison of phenotypes requires reasoning over multiple ontologies [26]. For many purposes it can be useful to extract modules from the union of CL with species-specific anatomy ontologies.

Taxonomic modularization requires a slightly different strategy. This was first proposed and formalized by Kusnierczyk [18], and later implemented in the GO [7]. For example, the GO states that lactation occurs only in Mammalia, allowing a module extraction tool to automatically generate a *Drosophila* subset that excludes this term. However, other requirements, such as generating labels specific to certain taxa remain unmet.

As a multi-species ontology that is integrated with multiple species-specific anatomy ontologies (AOs) (see lower half of Table 1), the CL has particular requirements here. There is overlap between the general terms in CL and the species-specific terms in these AOs, with the degree of overlap varying depending on the ontology.

For example, there is little overlap between plant and metazoan cell types, so it makes sense to manage these in separate ontologies. The Plant Ontology (PO), which combines cells and gross anatomy in a single ontology, is taking responsibility for plant cell types, leaving CL to focus on metazoa.

The situation is more complex when we consider the *Drosophila* Anatomy Ontology (FBbt). Managing all *Drosophila* cell types in CL would be difficult: this ontology has over 1,500 neuronal cell types, many of which are specific to this taxon, and this number is likely to grow. Representing these cell types in FBbt allows linkages between cell types and *Drosophila* gross anatomy to be maintained in a simple and logically coherent way. At the same time, we want to coordinate on a shared representation of core cell types such as “neuron” in the CL. We also want CL to have very specific cell types for mammals (note that the adult Mouse Anatomy ontology, MA, does not represent cell types). This tension between a shared general representation and individual specific representations creates challenges for ontology management. In addition, many users want to be able to obtain a single coherent ontology view of all cell types within a clade, or across all organisms, requiring intelligent combination of multiple ontologies.

The strategy thus far has been for CL to represent generalized cell types as far as possible, with taxonomic specificity indicated by constraints in the ontology, and for taxon-centric ontologies such as ZFA and FBbt to represent these cell types as they are instantiated in particular species, with OBO format “xrefs” (semantics-free cross-references) linking the two.

Towards an Integrated Tool

There is a lack of a single integrated tool that can fulfill all of these requirements. In an effort to redress this, we have specified a list of capabilities such a tool should have for working with the CL, and present initial progress towards the implementation of such a tool.

2 Cell Ontology Requirements

Axiom Rewriting Using Subsets

A class subset S is a collection of classes c^1, c^2, \dots, c^n taken from an ontology O . An ontology O

can be rewritten as an ontology O' such that O' contains no references to classes not in S , yet is still consistent with O . This process is colloquially known in GO as *slimming*. Note that we use the term ontology in the sense of any collection of axioms; this means that if we have a formalization of GO associations in OWL, we can use the same algorithm for mapping associations.

Note that the axioms in the target ontology need not be a subset of the axioms in the source ontology – some axiom rewriting may be required. Consider the case where X is a subclass of Y and Y is a subclass of *part_of* some Z , and $S = X, Z$, then a simple subsetting operation will lose the axioms connecting X to Z . The following procedure should be used to extract a subset S from ontology O :

- Create a target ontology O' that is an exact copy of the source ontology O
- Remove all axioms from O' where that axiom references a class not in S (i.e. all classes in the signature of the axiom must be in S).
- Reason over O to find all inferred axioms¹ A .
- For each axiom in A , add that axiom to O' , provided this is not redundant with anything in O' . An axiom is redundant if it exactly matches an existing axiom, or it is entailed by O' .

For OBO format (obof) ontologies, the ontology should first be converted to OWL, after which it can be converted back to obof; this ensures correct interpretation when implementing the above procedure.

If the source ontology contains equivalence axioms (*intersection_of* tags in obof) that reference a class not in the subset, this procedure will rewrite them as plain SubClassOf axioms (*is_a* or *relationship* tags in obof). This is the desired behavior, as writing the IDs but keeping equivalence axioms would result in incorrect inferences.

Ontology Property Subsets

Some ontologies use a large number of properties (relations), some of which may be organized in a hierarchy. For example, the

¹ Whilst strictly speaking “inferred axiom” is an oxymoron, the OWL literature uses “axiom” in place of “sentence” and frequently distinguishes between inferred and asserted axioms

FMA has many different relations, and distinguishes between 3 sub-properties of *part_of* (systemic, regional and constitutional). Sometimes it is desirable to map these to the generic relation.

Here we can specify an ontology *property subset*, excluding the sub-properties of *part_of*. Then when we use the procedure above, axioms are automatically “mapped up” to the generic relation.

For example, if the source ontology contains an axiom *X* subclass of *regional_part_of* some *Y*, and the regional part relation is a sub property of *part_of*, then a reasoner can infer *X* subclass of *part_of* some *Y*. If the property subset contains only the generic relation, then the target ontology would have only the latter axiom and not the former.

Annotation Axiom Rewriting

When constructing a union of the general Cell Ontology and a species-specific ontology such as FBbt, we are faced with a problem that the resulting ontology will result in classes with non-unique labels, since we will have both CL:0000540 (neuron) and FBbt:00005106 (neuron). One highly impractical solution is for each anatomy ontology to ensure their primary labels are globally unique – for example, FBbt:00005106 would be labeled “*Drosophila* neuron”. Another approach would be to merge selected class pairs as part of the process of creating the union – for example, merging FBbt:00005106 into CL:0000540. One must then decide how to deal with the axioms of the merged classes. If the axioms are combined it can generate problematic statements such as “every (generic) neuron is part of a *Drosophila* nervous system” – obviously false for a zebrafish Purkinje cell. The opposite approach, discarding axioms, loses potentially useful information.

The accepted solution is to create a *bridge ontology* connecting the ontologies, and include annotation assertions in this bridge ontology for multi-context labels. For example, the bridge ontology would assert that FBbt:00005106 would have an “OBO Foundry unique label” of “*Drosophila* neuron” or “neuron (*Drosophila*)” where the taxon is included in the label. This would only be necessary for taxon-specific subclasses of generic classes, but it may be simpler to apply this uniformly across the species-specific anatomy ontology.

The modularization procedure can then merge the generic Cell Ontology with the cell subsets of the species-centric anatomy ontologies and rewrite the primary label axiom to use the OBO Foundry label, adding an axiom annotation to the axiom stating the source of the rewriting.

Taxonomy Reasoning Based Module Extraction

Many ontologies, such as GO and CL, are intended to be applicable across taxa. This means that these ontologies typically contain modules that are useful to one community and not another; for example the class *mammary gland epithelial cell* in the CL would not be useful for gene expression queries for chicken. The taxonomic constraint strategy used by GO [7] has been adapted for UBERON and will be used for the CL, replacing the *sensu* designators that are currently in use.

One of the main use cases for taxon-based module extraction from the CL would be to provide modules that exclude non-taxonomically relevant classes. For example, in a generic cell ontology it is useful to have a generic “erythrocyte” class and two subclasses, depending on whether the erythrocyte is nucleate or enucleate. However, most species have one or the other form, so when creating a taxon module the irrelevant classes can be discarded. For example, for a mouse module, only “enucleate erythrocyte” is required. It may also be desirable to give this the label “erythrocyte” when used in a mouse context.

This kind of automated taxonomic extraction is possible, provided the ontology has enough axioms to support this. For the above example, the ontology would have to state that (1) “erythrocyte” is a the disjoint union of “enucleate erythrocyte” and “nucleate erythrocyte” (alternatively, this could be inferred if these classes are defined using GO) and (2) no mammal erythrocyte is a nucleate erythrocyte (i.e. a standard taxonomic constraint). The taxonomy ontology will tell us that mouse is a subclass of mammal. The module extraction procedure for a taxon *t* is then to add an axiom for every class *c* in the ontology, stating that *c* is in-taxon some *t*. We then eliminate any unsatisfiable classes, and merge equivalent classes, using the more generic label as primary.

Taxonomic Bridge Ontologies

Cross-species ontology integration can be assisted by means of *bridging axioms* – for example, ZFA:0009248 (neuron in ZFA) is a subclass of the generic CL:0000540 (neuron in CL).

Maintaining these bridging axioms explicitly can be difficult since the resulting ontology has a highly latticed structure, so an alternative approach is to use a feature specified in obo format 1.4 called “*xref macros*”. Here header directives can be used to indicate how xrefs for a particular ontology are to be translated. For example, use a *treat-all-xrefs-as-has-subclass* header directive in CL, all FMA xrefs in CL can be expanded to:

```
Class: FMA_54527
SubClassOf: CL_0000540
```

We can even make a stronger taxonomically-qualified equivalence axiom by including a *treat-all-xrefs-as-genus-differentia* directive together with appropriate IDs:

```
Class: FMA_54527
EquivalentTo:
```

```
CL_0000540 and
part_of some NCBITaxon_9606
```

i.e. any CL neuron in a human is equivalent to a FMA neuron. These header macros are applied on a per-ontology basis.

A modularization tool for CL should be capable of generating the logical axioms from the xrefs, and placing these in the requisite bridge ontologies, from where they can be subsequently merged.

Creation of Taxon-Union Importing Ontologies

We plan to publish modules that import subsets of both CL and external anatomy ontologies for different taxonomic clades. This is already possible to a certain extent with UBERON – figure 1 shows the OWL import chain for a pan-eukaryotic anatomy ontology which selectively imports pan-anatomy ontologies for different clades.

The modularization tool should be able to use taxon ontologies to dynamically build these importer ontologies.

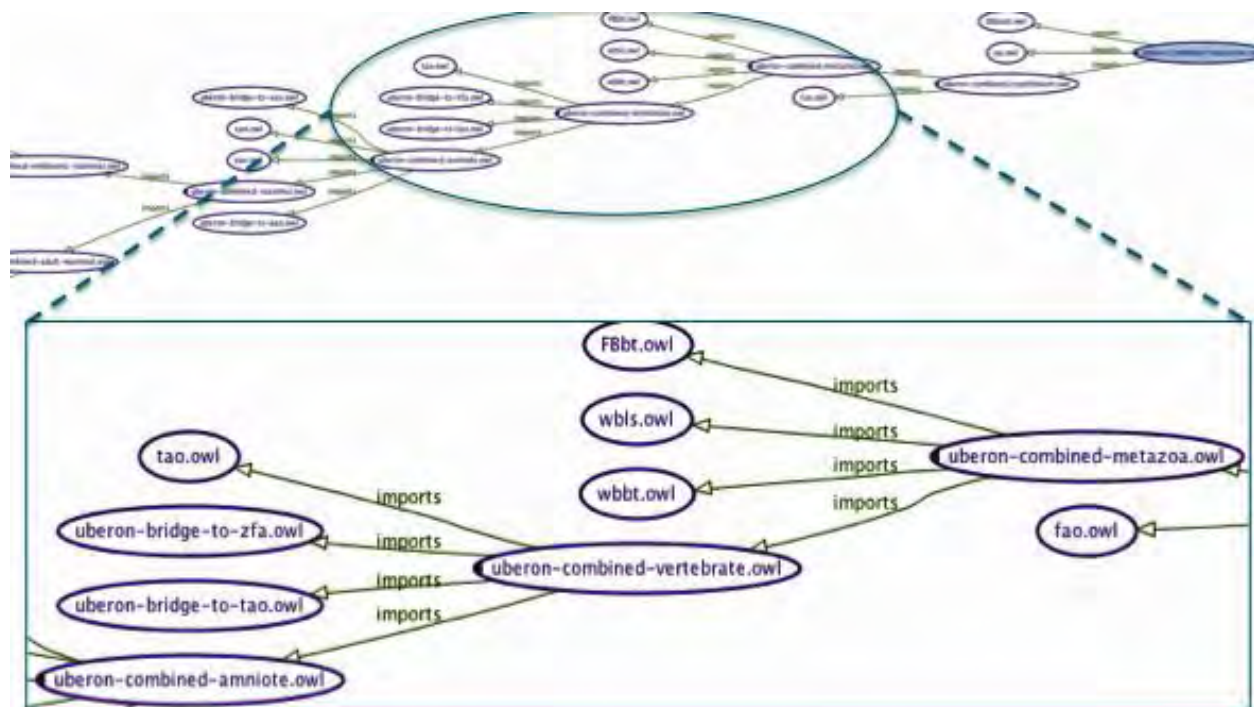


Figure 1. Import chain for uberon pan-eukaryote anatomy (<http://purl.obolibrary.org/obo/uber/mod/uberon-combined-eukaryote.owl>). This selectively imports bridge modules, species specific anatomy ontologies and recursively imports taxonomically more restricted import ontologies. The bridge modules are generated from xrefs stored in the main UBERON file. The zoomed area shows how the metazoa module imports the vertebrate module plus selected invertebrates, and the vertebrate module imports the amniote module plus selected anamniotes.

3 Implementation Progress

The CL modularization tool is being developed as part of the OWLTools library (<http://code.google.com/p/owltools/>), and will be released in the fall of 2011. OWLTools is layered on top of the OWLAPI [15], so it can take advantage of standard OWL reasoners and generic modularization tools. It also takes advantage of the new obo2owl implementation (<http://code.google.com/p/oboformat/>), and should thus be capable of working with ontologies whose source is either OBO format or OWL.

4 Conclusions

OWL modularization tools provide powerful and formally sound means of extracting modules from multiple ontologies that are amenable to reasoning. These tools would become even more useful for the bio-ontologies community if embedded in software that is aware of common metadata tags used in OBO ontologies and of taxonomic constraints. We have outlined some specific requirements for a generic tool that is currently being developed to perform these tasks.

References

1. M. Ashburner, C.J. Mungall, and S.E. Lewis. Ontologies for biologists: a community model for the annotation of genomic data. In *Cold Spring Harbor symposia on quantitative biology*, volume 68, pages 227–235, 2003.
2. Jonathan Bard, Seung Y Rhee, and Michael Ashburner. An ontology for cell types. *Genome Biol*, 6(2):R21, 2005.
3. M.C. Costanzo, M.S. Skrzypek, R. Nash, E. Wong, G. Binkley, S.R. Engel, B. Hitz, E.L. Hong, and J.M. Cherry. New mutant phenotype data curation system in the *Saccharomyces* Genome Database. *Database*, 2009(0), 2009.
4. M. Courtot, F. Gibson, A.L. Lister, J. Malone, D. Schober, R.R. Brinkman, and A. Ruttenberg. Mireot: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33, 2011.
5. W. M. Dahdul, J. G. Lundberg, P. E. Midford, J. P. Balhoff, H. Lapp, T. J. Vision, M. A. Haendel, M. Westerfield, and P. M. Mabee. The Teleost Anatomy Ontology: Anatomical Representation for the Genomics Age. *Syst Biol*, 59(4):369–383, 2010.
6. John Day-Richter, Midori A Harris, Melissa Haendel, Gene Ontology OBO-Edit Working Group, and Suzanna Lewis. OBO-Edit—an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200, Aug 2007.
7. J. Deegan, E. Dimmer, and C.J. Mungall. Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC bioinformatics*, 11(1):530, 2010.
8. R. A. Drysdale, M. A. Crosby, W. Gelbart, K. Campbell, D. Emmert, B. Matthews, S. Russo, A. Schroeder, F. Smutniak, P. Zhang, P. Zhou, M. Zytkevich, M. Ashburner, A. de Grey, R. Foulger, G. Millburn, D. Sutherland, C. Yamada, T. Kaufman, K. Matthews, A. DeAngelo, R. K. Cook, D. Gilbert, J. Goodman, G. Grumblin, H. Sheth, V. Strelets, G. Rubin, M. Gibson, N. Harris, S. Lewis, S. Misra, and S. Q. Shu. FlyBase: genes and gene models. *Nucleic Acids Res*, 33(Database issue):D390–5, 2005. 1362-4962 Journal Article.
9. G.V. Gkoutos, C. Mungall, S. Doelken, M. Ashburner, S. Lewis, J. Hancock, P. Schofield, S. Khler, and P.N. Robinson. Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)*, 2009.
10. M. Haendel, M. Wilson, C. Torniai, E. Segerdell, C. Shaffer, R. Frost, D. Bourges, J. Brownstein, and K. McInerney. Eagle-i: Making invisible resources, visible. *Journal of Biomolecular Techniques: JBT*, 21(3 Suppl):S64, 2010.
11. Melissa A. Haendel, Georgios G. Gkoutos, Suzanna E. Lewis, and Chris Mungall. Uberon: towards a comprehensive multi-species anatomy ontology, August 2009.
12. M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32 Database issue:D258–61, 2004. 1362-4962 Journal Article.

13. T.F. Hayamizu, M. Mangan, J.P. Corradi, J.A. Kadin, and M. Ringwald. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biology*, 6(3):R29, 2005.
14. D. P. Hill, J. A. Blake, J. E. Richardson, and M. Ringwald. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res*, 12(12):1982–91, 2002. 1088-9051 Journal Article.
15. M. Horridge. The OWL API: A Java API for Working with OWL 2 Ontologies. In *6th OWL Experiences and Directions Workshop (OWLED 2009)*, 2009.
16. P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, et al. Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6(7):388–397, 2005.
17. B. Konev, C. Lutz, D. Walther, and F. Wolter. Semantic modularity and module extraction in description logics. In *Proceeding of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 55–59. IOS Press, 2008.
18. Wacław Kusnierczyk. Taxonomy-based partitioning of the Gene Ontology. *Journal of Biomedical Informatics*, 41(2):282–292, 2008.
19. R.Y.N. Lee and P.W. Sternberg. Building a cell and anatomy ontology of *Caenorhabditis elegans*. *Comparative and Functional Genomics*, 4(1):121–126, 2003.
20. Terrence Meehan, Anna Maria Masci, Amina Abdulla, Lindsay Cowell, Judith Blake, Christopher Mungall, and Alexander Diehl. Logical development of the cell ontology. *BMC Bioinformatics*, 12(1):6, 2011.
21. Christopher J. Mungall, Michael Bada, Tanya Z. Berardini, Jennifer Deegan, Amelia Ireland, Midori A. Harris, David P. Hill, and Jane Lomax. Cross-Product Extensions of the Gene Ontology. *Journal of Biomedical Informatics*, 44(1):80 – 86, 2011. Ontologies for Clinical and Translational Research.
22. Darren A Natale, Cecilia N Arighi, Winona C Barker, Judith Blake, Ti-Cheng Chang, Zhangzhi Hu, Hongfang Liu, Barry Smith, and Cathy H Wu. Framework for a protein ontology. *BMC Bioinformatics*, 8 Suppl 9:S1, 2007.
23. Alan L. Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 121–128, Sanibel Island, FL, USA, 2003. ACM.
24. C. Rosse and J.L.V. Mejino. A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics* 36:478-500.
25. Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251– 1255, Nov 2007.
26. Nicole L Washington, Melissa A Haendel, Christopher J Mungall, Michael Ash-burner, Monte Westerfield, and Suzanna E. Lewis. Linking Human Diseases to Animal Models using Ontology-based Phenotype Annotation. *PLoS Biology*, 7(11), 2009.
27. Z. Xiang, M. Courtot, R.R. Brinkman, A. Ruttenberg, and Y. He. OntoFox: web-based support for ontology reuse. *BMC research notes*, 3(1):175, 2010.

Building the OBO Foundry – One Policy at a Time

Mélanie Courtot¹, Chris Mungall², Ryan R. Brinkman^{1,3}, Alan Ruttenberg⁴

¹BC Cancer Agency, Vancouver, BC, Canada

²Lawrence Berkeley National Laboratory, Berkeley, CA, USA

³Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

⁴University at Buffalo, NY, USA

Abstract. Policy drafting, discussion and implementation is not the most exciting or interesting thing to do when developing new resources. However, when trying to identify existing work that can be built upon in one's project, such policies are critical to allow interoperability and reliability. We describe some tools and guidelines developed under the OBO Foundry umbrella, and show how they help realize critical maintenance functions, increasing overall quality and sustainability of resources.

1 Introduction

With the increasing number of ontologies created in the biomedical domain, the ability to work with multiple resources is critical for developers. This allows developers to concentrate on new requirements rather than duplicate existing effort. However, in order to harmoniously build on several distinct bodies of work originating from different communities, guidelines should be established and followed. We present our experience taking part in the Open Biomedical Ontologies (OBO) Foundry [1] consortium. We briefly summarize some work that has been done under the OBO Foundry umbrella in defining a common ID policy¹, and a shared metadata set incorporated in the Information Artifact Ontology (IAO) [2]. Finally, we describe the issues related to a lack of a common deprecation policy, and propose a process for harmonizing expected behaviour across resources.

2 The OBO Foundry

The OBO Foundry is a set of ontologies which are intended to be interoperable, designed following a similar philosophy and implemented in accordance with a set of principles and guidelines. Authors of resources submitted to the OBO Foundry library² commit to working

together to increase quality of resources. As a result of that collaborative work, resources part of the OBO Foundry are orthogonal in scope (i.e., each resource describes a specific, non-overlapping domain) – and common policies are devised and followed. To increase interoperability, ontologies use a common upper ontology (Basic Formal Ontology (BFO)) [3] and a common set of relations (Relations Ontology (RO)) [4]. Policy adoption at the level of the OBO Foundry is done by decision of the OBO coordinators, a set of individuals whose task is to help build a community adhering to the OBO principles, and facilitate collaborations and cooperation between groups and resources.³

Common Unique Identifier Policy

The OBO foundry currently hosts resources under the OBO format [5] and the Web Ontology Language (OWL) [6] format, and aims at providing tools such as the OWLAPI mapping for OBO format⁴ to allow their interconversion. In order to do so, one key requirement is to rely on a common system to handle unique identifiers for entities. A policy, normative for Foundry resources, includes a Foundry-compliant Uniform Resource Identifier (URI) scheme, and rules to map from current OBO IDs and OBO legacy URIs towards them. Following a common ID policy allows URIs to be more reliable, and ensures they are unique within the Foundry consortium. It also helps

¹ The OBO Foundry policy is available at
<http://www.obofoundry.org/id-policy.shtml>

² <http://www.obofoundry.org/>

³ <http://obofoundry.org/coordination.shtml>

⁴ <http://code.google.com/p/oboformat/>

building tools relying on this ID scheme. For example, the Ontology of Biomedical Investigations (OBI) [7] developers do not deal with ID management when creating entities; rather a script is run pre-release to check and homogenize URIs for format and stability (e.g., was any URI deleted since the last release?). Another feature is to allow dereferencing and provide useful information to a user trying to resolve terms' URIs. The OntoBee browser⁵ displays a HTML page that provides human readable information on each term, such as label and textual definition, while the page source is RDF that can be machine-processed. Finally, the ID policy specifies versioning rules for ontology releases, effectively creating a version history for resources. By doing so, users are always free to access the latest published version and get the most up to date developments, or instead use a specifically dated release, and maintain stability of their own resource.

3 Improving Documentation by Sharing Metadata through the IAO

The IAO is an ontology of information entities, which aims at providing high-level blocks upon which specific resources can build upon. It describes classes such as *directive information entity*, which can for example be extended in a clinical-focused ontology by the *clinical guideline* subclass. As part of the IAO project, a distinct file defining common metadata properties⁶ has been created. This file can be imported independently of the “core” IAO, and used by any developer. The IAO common metadata set contributes to the realization of the principle of documenting ontologies within the OBO Foundry.

Other efforts already exist to formalize metadata, such as the Simple Knowledge Organization System (SKOS) [8] and the Dublin Core (DC) Metadata element set [9]. However, we found them not adequate for our usage. If we consider the case of `dc:creator`, its definition reads “An entity primarily responsible for making the resource.”, where

the resource is the resource described by the class bearing this property. For example, if we describe a book, the `dc:creator` property value is set to the name of the author of the book, and does not capture the name of the author of the book description, which is what we would aim at capturing with `iao:definition_editor`⁷. Similarly, the definition of `skos:definition` defines concepts, which is not suitable in our case.

Common and expected annotation properties, such as *definition* and *editor preferred term* are documented, and allow tool developers to rely on them to build their user interface. Other properties such as *definition source* or *definition editor* were created to store any references used in developing the definition and who did created the term. This allows resource consumers to go back and check on the origin of the term and what its intended meaning is, and/or contact the relevant individual should they need more clarification about its usage. Similarly, curators of the ontology can add *example of usage* and *editor note* to further clarify what the term denotes and what its intended usage is. Other slightly more complex properties have been designed to enable quality assessment of the terms. Namely, the *curation status specification* class provides a list of predefined instances (i.e., ‘*example to be eventually removed*’, ‘*metadata complete*’, ‘*organizational term*’, ‘*ready for release*’, ‘*metadata incomplete*’, ‘*uncurated*’, ‘*pending final vetting*’, ‘*to be replaced with external ontology term*’, ‘*requires discussion*’⁸) that can be used on each class to mark its degree of “readiness” and stability. Similarly, the class *obsolescence reason specification* offers a list of predefined values that can be used on obsoleted terms to give more information as to why that term was deprecated and indicate (in conjunction with for example an editor note) what the term replacement is. Finally, an *OBO Foundry unique label* annotation property (http://purl.obolibrary.org/obo/IAO_0000589), was recently added in the ontology-metadata file to allow disambiguation between

⁵ <http://www.ontobee.org/>

⁶ <http://purl.obolibrary.org/obo/iao/ontology-metadata.owl>

⁷ <http://dublincore.org/documents/dcmes-xml/>, section 2.4

⁸ <http://code.google.com/p/information-artifact-ontology/wiki/OntologyMetadata>

terms local to a resource when they are taken in the whole set of OBO Foundry ontologies. *OBO foundry unique labels* are automatically generated based on regular expressions provided by each ontology, when processed by the OBO package manager currently being written by the OBO Foundry custodians.

4 Maintaining Orthogonality through MIREOT

The OBO Foundry requires that newly created ontologies be orthogonal to resources already lodged within OBO. As a consequence, when in implementing a new resource, care should be taken to reuse work done in the context of other efforts where possible. Additionally, reusing terms from other resources allows developers to rely on the knowledge of domain experts who curated them and to dedicate more work time for novel terms *de novo*. Avoiding duplication of resources increases interoperability. A single URI is created per term, preventing the need for tedious mappings between terms with the same meaning in different resources.

When only few terms of interest are identified in external ontologies, those can be imported relying on the Minimum Information to Reference an External Ontology Term (MIREOT) guideline [10]. For example the Vaccine Ontology (VO) [11] defines the *vaccination* process as an “administering substance in vivo that involves in adding vaccine into a host (e.g., human, mouse) in vivo with the intend to invoke a protective immune response”, and the Adverse Events Reporting Ontology (AERO) [12] uses it as a synonym of the “immunization process” needed to define vaccine adverse events. The MIREOT mechanism provides a way to selectively import a term from a source ontology into a target resource, without the overhead of importing the whole external file. A more complete discussion pertaining to the trade-off of using MIREOT vs. other options, such as *modules* [13], is available in the MIREOT manuscript [10].

5 Deprecation Policy, an Unmet Need

Sometimes terms need to be retired as

ontologies evolve. The OBO Foundry doesn't currently formalize a standard deprecation policy, which leads to the problem of different policies within resources. As a general guideline, deprecated terms are not deleted from the ontology, as removing a term that has been used in the past can be confusing for users. Some discrepancies exist between the practice of the Gene Ontology (GO) [14] and OBI: in the GO [15], when terms are merged, one term effectively disappears from the ontology file and its identifier is maintained as an *alt_id* annotation property on the term it is merged with. By contrast in the OBI, one term is deprecated, and its *obsolescence reason specification* is set to “term merged”, with the addition of an editor note indicating the replacement term. As a consequence, tools such as MIREOT, developed in the context of the OBI, expect to find the URI of classes in their declaration (and not as a secondary ID). MIREOT scripts are therefore unable to retrieve the external information in the GO merging case, resulting in a loss of terms on the importing ontology side, such as recently happened with some Phenotypic Quality Ontology (PATO) terms [16]⁹. A common deprecation policy, following the example of what has been done regarding the ID policy, would help formalize expected behaviour, and guide tools developers. A review of the current reasons for obsolescence in the GO would be useful to perform to ensure adequacy between the instances defined by the IAO and the needs of the curators.

6 Evaluation

Most proposed policies have been adopted fairly recently, and evaluation is very preliminary. Although the relative costs and benefits could be difficult to quantify, a number of use cases illustrate the advantage of relying on numerical identifiers. For example, when choosing to use numerical IDs for terms, we know that some tooling issues will hinder adoption of those standards - nobody wants to type in OBI_0001234 when doing a SPARQL query. However, we believe that in the long term

⁹ http://sourceforge.net/mailarchive/forum.php?thre_ad_name=99D14FA3-9952-4C67-B892-41A8499A43C8\%40gmail.com&forum_name=obi-devel

(i) tooling issues will be resolved and (ii) using numerical IDs will be beneficial for maintenance of the resources and their necessary evolution. As illustration, see for example the recent threads mentioning how the (i) Protégé [17] team added a new menu “render by rdfs:label” to their interface¹⁰ and (ii) issues faced by the developer of GoodRelations [18] to rename some classes.¹¹

This paper is presented as a position paper/statement of interest. Our objective is to solicit feedback and interest from the community, and encourage participation in the development of current and future policies.

Acknowledgments

The authors’ work was partially supported by funding from the Public Health Agency of Canada / Canadian Institutes of Health Research Influenza Research Network (PCIRN), and the Michael Smith Foundation for Health Research. The authors wish to acknowledge people who contributed comments to the OBI deprecation policy: Suzanne Lewis, James Malone, Daniel Schober, Allyson Lister.

References

1. Ashburner M. Smith B., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L. J., Eilbeck K., Ireland A., Mungall C. J., Leontis N. OBI Consortium, Rocca-Serra P., Ruttenberg A., Sansone S. A., Scheuermann R. H., Shah N., Whetzel P. L., and Lewis S.. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.
2. The Information Artifact Ontology (IAO), <http://purl.obolibrary.org/obo/iao>.
3. The Basic Formal Ontology (BFO), <http://www.ifomis.org/bfo/>.
4. B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005.
5. The OBO Flat File Format Specification, version 1.2, http://www.geneontology.org/OLS.format.obo-1_2.shtml.
6. Web Ontology Language (OWL), <http://www.w3.org/2004/OWL/>.
7. OBI Ontology, <http://purl.obolibrary.org/obo/obi>.
8. Simple Knowledge Organization System (SKOS), <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>.
9. Dublin Core Metadata Element Set, <http://dublincore.org/documents/dces/>.
10. M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg. Mireot: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33, 2011.
11. The Vaccine Ontology, <http://www.violinet.org/vaccineontology/>.
12. The Adverse Event Reporting Ontology (AERO), <http://purl.obolibrary.org/obo/aero>.
13. Grau B.C., Horrocks I., Kazakov Y., and Sattler U. Extracting modules from ontologies: A logic-based approach. Proc. of the Third OWL Experiences and Directions Workshop, number 258 in CEUR, Innsbruck, Austria.
14. Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(90001):D258–D261, 2004.
15. GO editorial style guide, <http://www.geneontology.org/GO.usage.shtml>.
16. Phenotypic Quality Ontology (PATO), http://obofoundry.org/wiki/index.php/PATO:Main_Page.
17. The Protégé Ontology Editor and Knowledge Acquisition System, <http://protege.stanford.edu/>.
18. Martin Hepp. Goodrelations: An ontology for describing products and services offers on the web. In Aldo Gangemi and Jérôme Euzenat, editors, *Knowledge Engineering: Practice and Patterns*, volume 5268 of *Lecture Notes in Computer Science*, pages 329–346. Springer Berlin / Heidelberg, 2008.

¹⁰ <https://mailman.stanford.edu/pipermail/p4-feedback/2011-May/003889.html>

¹¹ <http://lists.w3.org/Archives/Public/public-lod/2011Apr/0278.html>

Towards a Body Fluids Ontology: A Unified Application Ontology for Basic and Translational Science

Jiye Ai¹, Mauricio Barcellos Almeida², André Queiroz de Andrade³,
Alan Ruttenberg⁴, David Tai Wai Wong¹, Barry Smith⁵

¹University of California at Los Angeles, School of Dentistry and Dental Research Institute, CA, USA

²Federal University of Minas Gerais, School of Information Science, Belo Horizonte, MG, Brazil

³Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

⁴State University of New York at Buffalo, Clinical and Translational Data Exchange, Buffalo, NY, USA

⁵State University of New York at Buffalo, Philosophy Department, Buffalo, NY, USA

Abstract. We describe the rationale for an application ontology covering the domain of human body fluids that is designed to facilitate representation, reuse, sharing and integration of diagnostic, physiological, and biochemical data. We briefly review the Blood Ontology (BLO), Saliva Ontology (SALO) and Kidney and Urinary Pathway Ontology (KUPO) initiatives. We discuss the methods employed in each, and address the project of using them as starting point for a unified body fluids ontology resource. We conclude with a description of how the body fluids ontology initiative may provide support to basic and translational science.

Keywords: body fluids, ontology, saliva, human blood.

1 Introduction

Body fluids are liquids that are excreted or secreted from and inside the bodies of organisms. We here focus on the case of human organisms, and provide a preliminary assay of the scope and purpose of an application

ontology covering the domain of body fluids in both healthy and diseased human organism.

Only a small fraction of the human body fluids have been included thus far in the Foundational Model of Anatomy (FMA) [1]. These are listed, with their definitions, in Table 1.

Body Fluid	Definition (FMA)
blood	Portion of body substance that consists of plasma and blood cells.
breast milk	Portion of secreted substance produced by the mammary gland.
chyme	Ingested food admixed with gastric secretions contained in the stomach.
endolymph	Transudate contained within the membranous labyrinth.
mucus	Portion of secreted substance produced by a mucous gland or goblet cell.
perilymph	Transudate contained in the osseous labyrinth outside the membranous labyrinth.
plasma	Body substance in liquid state contained in the lumen of arterial and venous trees, blood capillary and the cardiac chambers; constitutes the liquid phase of blood.
semen	Portion of body fluid suspension that consists of spermatozoa and seminal plasma.
sweat	Secretion produced by a sweat gland.
seminal fluid	Portion of secreted substance produced prostatic glands, bulbourethral glands or the seminal vesicles.
serum	Body substance derived from plasma by the elimination of fibrinogen.
tear	Portion of secreted substance produced by the lacrimal gland.
urine	Excretion in liquid state processed by the kidney.

Table 1. List of body fluid types currently represented in FMA

Further human body fluid types which we have identified in our researches thus far, and which will be submitted to the FMA for inclusion in due course, include:

- bile
- aqueous humour
- cerebrospinal fluid
- cerumen
- colloidal body substance
- deferent duct fluid
- epididymal duct fluid
- esophageal secretion
- follicular fluid
- gastric juice
- gingival fluids
- intestinal secretion
- intraocular fluid
- lymph
- epithelial lining fluid
- lung lining fluid
- menstrual fluid
- pancreatic juice
- renal filtrate
- rete testis fluid
- saliva
- sebum
- seminiferous tubule fluid
- serous fluid
- synovial fluid
- tissue fluid
- vaginal lubrication
- vitreous humour

Body fluids in a broader sense include also fluids such as liquid feces which exist in the organism in a state where they are dissolved in water. They include also fluids that result from procedures such as bronchial lavage. However, because the FMA deals only with body fluids present in the 'canonical' human body, it does not include terms representing fluids which arise in cases of disease and in the performance of clinical procedures.

There is a continuous flow of body fluids throughout the body. They function as vehicles to carry oxygen, nutrients, waste, hormones and other signal molecules and immune sensors and effectors between the body's different compartments. Body fluids serve also as transporters for pharmaceutical substances. Approximately 60 percent of the human body consists of fluids. Portions of fluid within the interior of the cell are called intracellular

fluids. All other fluids are extracellular, and it is these that we focus on here, and primarily on those extracellular fluids that are of value for diagnostic purposes.

Body fluids are present in the body in various combinations and in various proportions. An excess or shortage of a given body fluid in a given compartment or conduit can be a symptom or a cause of disease. Many diseases can affect body fluids in their turn, and the latter can thus serve as a diagnostic indicator of the former. This holds for some cancers [2, 3], kidney diseases [4], inflammatory diseases [5], and metabolic diseases [6]. Certain body fluids, above all blood, urine and saliva, provide advantages with regard to disease diagnosis and prognosis, primarily due to low invasiveness, minimal cost, and easy sample collection and processing [7, 8].

With the recent advances in biomedical technology, there has been an escalating need for formal tools that can facilitate effective and efficient representation, reuse, sharing and integration of diagnostic data. Scientists are increasingly recognizing the value of ontologies in this connection. An ontology provides a controlled vocabulary that can be shared by investigators in different fields, who can draw on the ontology's logical definitions to ensure that terms are used with common meanings. In addition the ontology can support quality control in data entry and allow algorithmic reasoning on data annotated using its terms.

The ontology-based approach will function successfully, however, only if the ontologies themselves are developed in tandem with each other in such a way as to ensure cross-domain consistency and to eliminate the sorts of redundancy in vocabulary creation which have traditionally arisen where domains overlap. To this end, a distinction needs to be drawn between reference and application ontologies [9]. The former correspond in medicine to the basic biomedical sciences such as anatomy and physiology, the latter to clinical specialisms and sub-specialisms, for example to pediatric surgery or radiation oncology. Just as the clinical specialisms draw on the content and results of the basic sciences, so application ontologies will need to draw on more basic feeder ontologies such as the aforementioned FMA, and others such as the Chemical Entities

of Biological Interest (ChEBI) [10], Protein Ontology (PRO) [11], Gene Ontology (GO) [12] and the Cell-Type Ontology (CT) [13].

Our proposed application ontology BFLO is designed to meet the need for terminology support in the domain of research on bio-fluids. As an application ontology, it will be built primarily out of terms deriving from other more foundational ontologies and terminologies. We shall draw most importantly on the FMA as our overarching anatomy framework [1]. There a *Portion of body fluid* is defined as:

A portion of body substance that consists of a mixture of fluid, solutes and particles.

FMA: *Portion of body substance* is defined in turn as:

Material anatomical entity in a gaseous, liquid, semisolid or solid state, with or without the admixture of cells and biological macromolecules; produced by anatomical structures or derived from inhaled and ingested substances that have been modified by anatomical structures.

All terms representing types of body fluid in our ontology will be treated as children of FMA: *Portion of body fluid*.

We will draw on other resources, including the SNOMED CT vocabulary of clinical terms, which includes terms relevant to the domain of body fluids such as *Body fluid (substance)*, *Origin of fluid (attribute)*, and *Body fluid retention (disorder)*. Body fluid (substance) is asserted in the SNOMED CT concept hierarchy to be both a Body substance and Liquid substance.

Apart from the FMA, the most important ontologies employed by the BFLO in its current alpha version are the Blood Ontology (BLO), the SALO Ontology (SALO) the Kidney and Urinary Pathway Ontology (KUPO).

1.1 Blood Ontology (BLO)

The BLO (<http://mbaserver.eci.ufmg.br/BLO-wiki/>) [14] is a controlled vocabulary designed for use in annotating and organizing data about blood, including data pertaining to:

- blood transfusions (for example, donation process control)

- hematology (for example, immunologic basis)
- blood derivative products (for example, frozen plasma)
- the content of regulatory documentation (for example, regulations under the Food and Drug Administration)
- the associated regulatory processes (for example tests of blood quality).

BLO is being created to serve the exploration and aggregation of information relevant to scientific research and to human blood manipulation. BLO is constructed on the basis of well-founded ontological principles in such a way that it can support interoperability with OBO Foundry ontologies such as the GO and the PRO. BLO corresponds to a set of interrelated ontologies, each addressing a group of relevant issues in the field of hematology and blood transfusion.

As concerns terms for diagnostic processes, BLO will rely on publications reporting the results of research on blood-transmitted diseases, for example, HIV-1/2, hepatitis B and C, Chagas Disease, and syphilis. In addition, the ontology will draw on terms used in research on hemophilias, Von-Willebrand disease and Sickle Cell Anemias, and on HTLV.

1.2 Saliva Ontology (SALO)

The SALO [15] is a consensus-based controlled vocabulary of terms and relations dedicated to the salivaomics domain and to saliva-related diagnostics. Like BLO, SALO follows the principles of the OBO Foundry.

The protein terms in the SALO are derived from the corresponding sections of the PRO (<http://purl.org/obo/owl/PRO>). The SALO is a component of an ontology-based framework for a Salivaomics Knowledge Base (SKB; <http://www.skb.ucla.edu/> website) that is a data repository, management system and web resource constructed to support human salivary proteomics, transcriptomics, miRNA, metabolomics and microbiome research currently being assembled in the Dental Research Institute at the UCLA School of Dentistry [15].

1.3 Kidney and Urinary Pathway Ontology (KUPO)

The KUPO (<http://www.e-lico.eu/public/kupo/kupo.owl>) [16] is an ontology that describes kidney and urinary anatomy, the cells in the associated organs and tissue, and the gene products in those cells and their functional attributes and cellular components and associated pathologies. It also contains terms relating to transcriptomics and proteomics experiments in the KUPO domain. Fluid-related terms in KUPO include: urea transport, small muscle layer of renal vein, rennin secretion into blood stream, and disorder of ureter.

2 Method

All body fluids share certain features in common, including relations to biological process represented in the GO, to proteins represented in the PRO, and to cell types represented in the CT. Given the large variety of different types of body fluid, however, these common features must be specialized to each individual case. In developing BFLO we accordingly employ a novel method for ontology development, which we call the method of generalization and specialization.

In the simplest case, BLO and SALO are taken as an input, and corresponding terms in the two ontologies, such as *portion of blood* and *portion of saliva* are aligned in light of their common Basic Formal Ontology (BFO) [17] category (<http://www.ifomis.org/bfo/>). The aligned terms are then joined to form a new ontology term as output, for example: *portion of blood*, *portion of saliva* → *portion of body fluid*. The MIREOT [9] guidelines are followed at every stage when terms are imported into BFLO from other ontologies.

3 Results and Discussion

Figure 1 displays a fragment of the BFLO illustrating fundamental entities such as water, protein, DNA, RNA, cell, organ, etc.

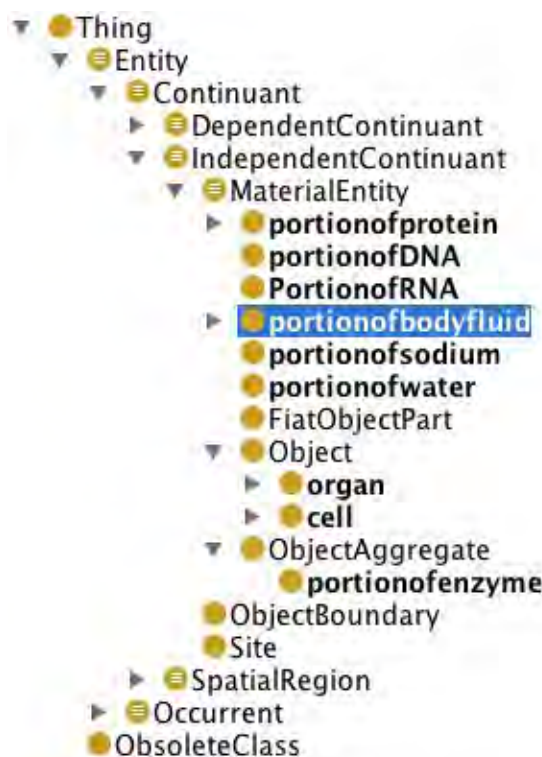


Figure 1. Fragment of BFLO

Application ontologies may be created either because reference ontologies are too large to be of effective service in specific, narrowly focused projects, or because they are too narrow in scope. We might, for example, be interested in data associated with physiological models of body fluid exchange [18], or in data which results from comparing the utility of different body fluids (for example saliva, blood and urine) for diagnostic purposes in relation to multiple different sorts of diseases. An application ontology such as BFLO can then help us to focus on the relevant content of the reference ontologies and to conjoin the corresponding fragments together in a way that helps us to address cross-domain issues.

Another set of examples of potential uses of BFLO concern comparisons of the diagnostic value of different body substances where the results of tests employing one substance point to the need for tests using some other substance.

To build an adequate representation of a biological phenomenon we need precise information about the biological components involved on several different levels of granularity. Proteomics researchers [19] have mapped the similarities and differences in

protein composition in plasma and saliva samples. They have shown, for instance, that immunoglobulins present in saliva and plasma overlap in a way that suggests leakage from plasma into saliva. Given the convenience of saliva sampling such leakage could provide the possibility of more expedient testing for antibodies. At the same time, it may be that we can use information about known biomarkers in blood to make inferences to the presence of as yet unknown biomarkers in saliva [20]. Integration of blood and saliva data could thus be advanced in useful ways through the creation of the unified BFLO framework. The representation would need to take into account also the fact that some proteins have different behaviors in special conditions (such as diseases), and thus proper functional annotation is essential if an ontology-based representation of data is to support prediction.

4 Conclusion and Future Directions

The full understanding of the physiology and of the body requires the use of data relating not merely to body fluids taken singly, but also in combination with other body fluids, and of course with other anatomical entities.

The techniques and applications sketched in the foregoing are just the first steps towards a truly useful BFLO. Close coordination with the OBO Foundry, consolidation of a common framework, and exploration of potential collaborations with KUPO, as well as expansion to other representative fluids, are the next steps in which we intend to move forward.

The growth of research on body fluids is raising important challenges for the field of biomedical ontologies, bringing the demand for a resource that can accelerate the consistent representation, organization and manipulation of body fluid data. We accordingly believe that a resource like the proposed BFLO has the opportunity to advance both basic and translational science.

Acknowledgments

This work was supported by US National Institutes of Health grants U01DE016275, U01DE017790 and U54HG004028.

References

1. Rosse, C., Mejino, J.L., Jr.: A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform* 36, 478-500 (2003)
2. Gao, K., Zhou, H., Zhang, L., Lee, J.W., Zhou, Q., Hu, S., Wolinsky, L.E., Farrell, J., Eibl, G., Wong, D.T.: Systemic disease-induced salivary biomarker profiles in mouse models of melanoma and non-small cell lung cancer. *PLoS One* 4, e5875 (2009)
3. Wulfkühle, J.D., Liotta, L.A., Petricoin, E.F.: Proteomic applications for the early detection of cancer. *Nat Rev Cancer* 3, 267-275 (2003)
4. Israni AK, e.a.: Laboratory assessment of kidney disease: Clearance, urinalysis, and kidney biopsy. Saunders Elsevier, (2008)
5. Young, B., Gleeson, M., Cripps, A.W.: C-reactive protein: a critical review. *Pathology* 23, 118-124 (1991)
6. Burton, B.K.: Inborn errors of metabolism in infancy: a guide to diagnosis. *Pediatrics* 102, E69 (1998)
7. Greer, J.P.e.a.: Wintrobe's clinical hematology. Lippincott Williams & Wilkins, (2008)
8. Veenstra, T.D., Conrads, T.P., Hood, B.L., Avellino, A.M., Ellenbogen, R.G., Morrison, R.S.: Biomarkers: mining the biofluid proteome. *Mol Cell Proteomics* 4, 409-418 (2005)
9. Courtot, M., Gibson, F., Lister, A.L., Malone, J., Schober, D., Brinkman, R.R., Ruttenberg, A.: MIREOT: The minimum information to reference an external ontology term. *Appl. Ontol.* 6, 23-33
10. de Matos, P., Alcantara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C.: Chemical Entities of Biological Interest: an update. *Nucleic Acids Res* 38, D249-254 (2010)
11. Natale, D.A., Arighi, C.N., Barker, W.C., Blake, J., Chang, T.C., Hu, Z., Liu, H., Smith, B., Wu, C.H.: Framework for a protein ontology. *BMC Bioinformatics* 8 Suppl 9, S1 (2007)
12. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29 (2000)
13. Bard, J., Rhee, S.Y., Ashburner, M.: An ontology for cell types. *Genome Biol* 6, R21 (2005)

14. Almeida MB., S.B., Proietti ABC. Coelho KC.: The Blood Ontology: organizing the information in the domain of the human blood.
15. Ai, J., Smith, B., Wong, D.T.: Saliva Ontology: an ontology-based framework for a Salivaomics Knowledge Base. *BMC Bioinformatics* 11, 302 (2010)
16. Simon Jupp, J.K., Joost Schanstra and Robert Steven.: Developing a Kidney and Urinary Pathway Knowledge Base. *Bio-ontologies SIG* 2010, Boston, USA (2010)
17. Grenon, P., Smith, B., Goldberg, L.: Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform* 102, 20-38 (2004)
18. Thomas, S.R., Baconnier, P., Fontecave, J., Francoise, J.P., Guillaud, F., Hannaert, P., Hernandez, A., Le Rolle, V., Maziere, P., Tahi, F., White, R.J.: SAPHIR: a physiome core model of body fluid homeostasis and blood pressure regulation. *Philos Transact A Math Phys Eng Sci* 366, 3175-3197 (2008)
19. Yan, W., Apweiler, R., Balgley, B.M., Boontheung, P., Bundy, J.L., Cargile, B.J., Cole, S., Fang, X., Gonzalez-Begne, M., Griffin, T.J., Hagen, F., Hu, S., Wolinsky, L.E., Lee, C.S., Malamud, D., Melvin, J.E., Menon, R., Mueller, M., Qiao, R., Rhodus, N.L., Sevinsky, J.R., States, D., Stephenson, J.L., Than, S., Yates, J.R., Yu, W., Xie, H., Xie, Y., Omenn, G.S., Loo, J.A., Wong, D.T.: Systematic comparison of the human saliva and plasma proteomes. *Proteomics Clin Appl* 3, 116-134 (2009)
20. Spielmann, N., Wong, D.T.: Saliva: diagnostics and therapeutic perspectives. *Oral Dis* 17, 345-354 (2011)

Connecting Ontologies for the Representation of Biological Pathways

Anna Maria Masci¹, Mikhail Levin², Alan Ruttenberg^{3,*}, Lindsay G. Cowell^{2,*}

¹Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA

²Department of Clinical Sciences, University of Texas Southwestern Medical Center at Dallas, TX, USA

³School of Dental Medicine, University at Buffalo, NY, USA and Creative Commons, Mountain View, CA, USA

*These authors contributed equally

Abstract. Significant effort has been put into the creation of a multitude of large, publicly available pathway databases. Most make their content available in at least one of several standard representation formats, but there are limitations to existing pathway representation formats, including underutilization of a common set of biomedical ontologies. To address this limitation, we developed an approach to representing biological pathways that relies on the use of ontologies from the Open Biomedical Ontologies (OBO) Foundry, including the Relation Ontology (RO), and adheres to the logical principles of ontology development advocated by the Foundry. To demonstrate the utility of this representation approach, we have curated comprehensive pathway representations for the signal transduction pathways initiated by seven of the mouse Toll-like receptors (TLR). Current efforts include the development of approaches for utilizing these representations for pathway analysis.

Keywords: OBO Foundry, ontology, signal transduction pathway, semantic web, OWL

1 Introduction

Biological pathways are central to biology. As a consequence, significant effort has been put into the creation of a multitude of large, publicly available pathway databases, for example Reactome (<http://www.reactome.org>), Kyoto Encyclopedia of Genes and Genomes (KEGG) (<http://www.genome.jp/kegg/>), and Pathway Commons (<http://www.pathwaycommons.org/pc/>). These databases provide tremendous value to the community by making vast quantities of pathway information available both for download and for web browsing.

While each of the pathway databases has its own internal representation, most make their content available in at least one of several standard representation formats, such as BioPax (<http://www.biopax.org>). The availability of pathway data in standard, machine-readable formats is extremely important given the frequent need to integrate pathway data obtained from different databases and to incorporate the data into custom-written bioinformatics algorithms.

Despite the widespread use of standard pathway representation formats, there are

limitations on the extent to which pathway representations can be integrated and jointly analyzed, arising from underutilization of a common set of biomedical ontologies and the lack of a formalism for their use. Some pathway databases, such as Reactome, utilize common ontologies like the Gene Ontology (GO) (<http://www.geneontology.org/>) to annotate some (but not all) of their pathway events. Other pathway databases, however, have independently developed their own ontologies. The use of common ontologies for pathway annotation is beneficial, not just for supporting interoperability between different pathway databases, but also for supporting interoperability with other information resources. For example, one can easily use the GO annotation of a Reactome event to query the GO database for a list of genes annotated with the GO term.

Even when common ontologies are used to annotate the molecules or events in a pathway, there are still difficulties in interpreting such pathway annotations, as the relation of a GO term to an annotated event is not clearly defined, and detailed domain knowledge is often needed to discern the exact connection in each case.

Finally, there are limitations on the extent to which information encoded in the ontologies can be brought to bear on pathway analysis. These limitations result from the fact that the common pathway representation formats use different semantics and logical formalisms from those used in common biomedical ontologies.

To address these limitations, we developed an approach to representing biological pathways that relies on the use of ontologies from the Open Biomedical Ontologies (OBO) Foundry [1], including the Relation Ontology (RO) [2], and adheres to the logical principles of ontology development advocated by the Foundry. Key features of our approach include (i) the use of ontology terms to designate each entity in a pathway, rather than their use solely as annotations, (ii) the use of RO relations to structure the connections among the entities in the pathway, and (iii) use of the same logical formalism as is used in developing OBO Foundry ontologies. We see multiple benefits to this approach. In particular, this approach significantly reduces the ambiguity that can exist between different pathway representations regarding the named entities, and it facilitates use of the taxonomic, partonomic, and other relations in the ontologies for querying and reasoning over pathways and during pathway analysis.

To demonstrate the utility of this representation approach, we have curated comprehensive pathway representations for the signal transduction pathways initiated by seven of the mouse Toll-like receptors (TLR). We have integrated a total of 379 terms from four ontologies, the three ontologies of the GO and the Protein Ontology (PRO) [3]. The resulting representation includes 290 material entities and 171 processes, with over 2000 relationships between them. Current work includes incorporation of terms from additional ontologies, such as the Cell Ontology (CL) [4], integration of the pathway representations with extended portions of the relevant ontologies, and application of the integrated resource to pathway analysis.

2 Methods

2.1 Representation Approach

To develop an ontology-based approach to representing biological pathways, we first

identified the types of entities to be included and the ontologies that have these types in their domains. We next identified the types of relationships that exist between the relevant entities and an appropriate set of formally defined relations to capture them. Finally, we developed a formal structure for representing pathways by specifying a set of high-level triples based on these relations.

We take as our case study the representation of signal transduction pathways which necessitates inclusion of the following types of entities: cell surface receptors, the cells on which the receptors are expressed, the molecules that participate in signaling events, the cellular locations in which these molecules can be found, and the protein domains each of the proteins has as part. In addition, we include the processes that unfold in the context of signaling, and the functions each of the entities carries out during signaling.

We selected OBO Foundry ontologies as the source for terms primarily because they are developed in a coordinated fashion, which facilitates their integration. In particular, they are developed as orthogonal (non-overlapping) ontologies, and they use a common set of relations, those from RO, applied in a consistent manner both within and between ontologies. The RO relations are formally defined which facilitates their use for computation, bringing additional benefits [2]. Finally, OBO Foundry ontologies are widely used, which facilitates integration of additional resources.

We import terms from OBO Foundry ontologies as follows:

- i) terms for proteins from the Protein Ontology (PRO) [3],
- ii) terms for molecular complexes from PRO and the Gene Ontology (GO) Cellular Component ontology (GOCC) [6],
- iii) terms for cells from the Cell Ontology (CL) [4, 7, 8],
- iv) terms from GOCC for cellular components that are the locations of molecules,
- v) terms for functions from the GO Molecular Function ontology (GOMF) [6],
- vi) terms for processes from the GO Biological Process (GOBP) ontology [6].

There is currently no OBO Foundry ontology for protein domains. Thus, protein domain terms are imported from Pfam [5],

which we interpret as referring to parts of proteins (i.e. material entities).

2.2 Representation of Toll-like Receptor Pathways

Creation of the ontology-based representation of TLR pathways involved the following steps. We first created a spreadsheet template to facilitate manual curation of the relevant information into the formal framework. The spreadsheet was designed to provide an intuitive organizational structure for biologist domain experts, to ease the process of entering information into the spreadsheet, and to facilitate the automated translation of the spreadsheet into a computable ontology representation format, such as OBO (<http://oboedit.org>) or the Web Ontology Language (OWL) (<http://www.w3.org/2007/OWL>). The template is set up such that each pathway is curated into a single spreadsheet comprised of two main parts, one part containing a list of all entities named in the pathway and one part containing a list of asserted relations that when taken together specify the participants in and structure of the pathway. The list of named entities includes, for each entity, a handle used to refer to the entity within the spreadsheet, the corresponding ontology term, and the unique identifier associated with the term in the OBO Foundry ontology that is the source for the term. The approach to representation is process-centric, and this is reflected in the list of asserted relations, the bulk of which relate a process to the molecules that participate in the process, as described below.

To curate relevant pathway information into spreadsheets, a domain expert reviewed the primary literature to obtain information for pathways initiated by receptors formed from TLR2, TLR3, TLR4, TLR5, TLR7, TLR8, and TLR9. For each entity included in the representation, the curator searched each of the ontologies listed above to identify the most appropriate term and obtain its unique identifier. The appropriateness of a term was determined by reading its definition as well as the definition of nearby terms (e.g. parent, child, sibling terms). When no appropriate term was available, a term request was made to the appropriate ontology. For each term, the name of the source ontology, the term label from the source ontology, and the term's

unique identifier from the source ontology were recorded in the templated spreadsheet.

Scripts to translate the templated spreadsheets into OWL 2 were written in common lisp (ABCL) (<http://common-lisp.net/project/armedbear/>), calling Java libraries (e.g. APACHE-POI and OWL-API). Imported terms were included using the Minimum Information to Reference an External Ontology Term procedure (<http://obi-ontology.org/page/MIREOT>).

3 Results

The foundation of our approach to representing biological pathways is the use of terms from OBO Foundry ontologies to name pathway entities, and the assertion of RO type-level relations between the entities to specify the pathway's structure (Figure 1). RO type-level relations (relations between classes) are defined in terms of RO instance-level relations (relations between individuals), and most are defined with an **all-some** structure. Thus, where capital letters indicate types (e.g. A, B), lower case letters indicate individuals (e.g. a, b), and R and R^* are type- and instance-level relations, respectively, the assertion $A\ R\ B$ is interpreted as follows: for **all** individuals a of type A , there exists **some** individual b of type B such that $a\ R^*\ b$. For example, 'nucleus *part_of* cell' is interpreted as: for any individual nucleus n , there exists some individual cell c such that $n\ part_of^*\ c$. We have also used a relation submitted for inclusion to RO (*realizes*) which is defined with an **all-only** structure interpreted as: for all individuals a of type A , if $a\ R^*\ b$ then $b\ is_a\ B$. Phrased another way: only individuals of type B can stand in relation R^* to individuals of type A . Triples such as $A\ R\ B$ are translated to OWL class expressions that encode the intended interpretation.

We formed a set of high-level triples that specify the types of assertions used in our approach to pathway representation. Each high-level triple specifies the RO type-level relation and the types of entities it joins. Specific versions of these triples are used to build each specific pathway representation. In the description below, RO relations are shown in italics, and terms referring to types of entities include a subscript indicating the

ontology from which the term was taken. For example, in the high-level triple templates, $\langle \text{protein} \rangle_{\text{PRO}}$ indicates that a term for a type of protein is taken from PRO. In a specific triple, ‘TLR4_{PRO}’ indicates that the term ‘TLR4’ is taken from PRO. In our representation the unique identifier from the source ontology identifies all terms, but for ease of presentation we use labels in what follows.

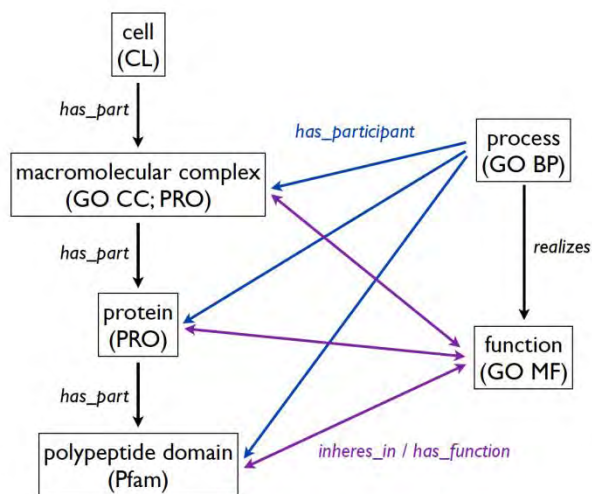


Figure 1: Ontology-based representation biological pathways.

Each box corresponds to a type of entity. Shown in parentheses are abbreviated names for the ontologies from which terms for entities of that type are imported. Arrows between boxes represent relations between the entity types. Blue arrows represent the *has_participant* relation. Purple arrows represent the *inheres_in* and *has_function* relations. Black arrows are as labeled.

3.1 Relations between Independent Continuants

The RO relation *has_part* as defined between types of independent continuants [2] is used to relate macromolecular complexes to their component proteins and to relate proteins to protein domains using the assertions

```

<macromolecular complex>GOCC/PRO1 has_part
<protein>PRO
<protein>PRO has_part <polypeptide domain>Pfam2.

```

¹ The term ‘macromolecular complex’ is a GOCC term. Terms for specific complexes are taken from either GOCC or PRO.

² The term ‘polypeptide domain’ is a Sequence Ontology term (<http://www.sequenceontology.org/>). Terms for specific types of domains are taken from Pfam via a prototype translation to OWL available on request.

Thus, we make no distinction between the type of part-whole relationship that obtains between complexes and their components and that which obtains between proteins and their domains. The RO *has_part* relation is defined with sufficient generality that it holds in both cases.

The TLR4 pathway representation includes these specific triples using the *has_part* relation:

```

TLR4:MD2PRO has_part TLR4PRO
TLR4:MD2PRO has_part MD2PRO
TLRPRO has_part TIR domainPfam.

```

These triples assert that TLR4:MD2 protein complexes have the proteins TLR4 and MD2 as part, and TLR proteins have TIR domains as part. Note that macromolecular complexes may have non-protein parts, which we do not currently specify.

To assert the relationship between cell types and their cell surface receptors (which may be proteins or complexes), we use the *has_plasma_membrane_part* relation used in the CL:

```

<cell>CL has_plasma_membrane_part
<macromolecular complex>GOCC/PRO
<cell>CL has_plasma_membrane_part
<protein>PRO.

```

For example,

```

dermal dendritic cellCL
has_plasma_membrane_part TLR4PRO.

```

has_plasma_membrane_part is defined in terms of the RO *has_part* relation and the GOCC term ‘plasma membrane’ [10] and is currently a candidate relation submitted to the RO.

3.2 Relations Between Independent Continuants and Processes

Fundamentally, pathways are collections of interconnected processes, linked through the requirement of one process for a participant produced by another process. Thus, relations between processes and their participants are central to our representation approach. We use the RO relation *has_participant* [2] and a proposed subrelation *has_output_participant* (http://www.berkeleybop.org/ontologies/obo-all/ro_proposed/ro_proposed.obo.html) to relate processes to the molecules that participate in them and are produced by them, creating assertions of the form

```

<process>GOBP has_participant <macromolecular
complex>GOCC/PRO

```

<process>GOBP *has_participant* <protein>PRO
 <process>GOBP *has_output_participant*
 <macromolecular complex>GOCC/PRO
 <process>GOBP *has_output_participant*
 <protein>PRO.

For example,

TLR4:MD2 complex assemblyGOBP
has_participant TLR4PRO
 TLR4:MD2 complex assemblyGOBP
has_participant MD2PRO
 TLR4:MD2 complex assemblyGOBP
has_output_participant TLR4:MD2PRO

which represents participation of TLR4 and MD2 in a process by which the TLR4:MD2 complex is formed.

To distinguish types of process participants, we relate the participants in a process to the functions they manifest in that process. To do so, we utilize relations submitted to RO and defined on the basis of the treatment of functions in the Basic Formal Ontology (BFO), the upper-level ontology for OBO Foundry ontologies. According to this treatment, functions are dispositions to participate in processes that belong to independent continuants and are manifested, or realized, when a continuant participates in a process of the relevant type. Thus, we have the following set of triples relating proteins to functions

<protein>PRO *has_function* <function>GOMF
 <function>GOMF *inheres_in* <protein>PRO

along with similar triples for protein domains and macromolecular complexes, and these triples relating functions and processes

<function>GOMF *realized_in* <process>GOBP
 <process>GOBP *realizes* <function>GOMF.

For example, the triple

TIR domainPfam *has_function* TIR domain
 bindingGOMF

asserts that TIR domains are capable of binding to other TIR domains. Similarly,

phosphorylationGOBP *realizes* kinase activityGOMF

asserts that phosphorylation processes are processes in which kinase functions are realized.

Under our approach, the full specification of a process involves assertions that combine the *realizes* relation with the *inheres_in* relation. For example, phosphorylation processes in which dual specificity mitogen-activated protein kinase kinase 3 serves as the kinase have the assertion

<process>GOBP *realizes* (kinase activityGOMF AND
 (*inheres_in* dual specificity mitogen-activated
 protein kinase kinase 3PRO)).

3.3 Relations between Processes

The RO *has_part* relation as defined between types of occurrents [2] is used to relate complex processes or sets of processes to their component processes:

<process>GOBP *has_part* <process>GOBP.

For example,

TLR4 signaling pathwayGOBP *has_part* I-kappaB
 phosphorylationGOBP.

Note that we do not specify any order to the processes in a pathway. These can be inferred from the participant assertions.

4 Discussion

We have developed an ontology-based approach to the representation of biological pathways with the goal of enhancing interoperability among pathway representations as well as between pathway and other information resources. Key features of our approach include (i) the use of terms from OBO Foundry ontologies to designate each entity in a pathway, rather than just as annotations, (ii) the use of RO relations to structure the pathway, and (iii) use of the same logical formalism as is used in developing OBO Foundry ontologies.

We anticipate several benefits from this approach. The use of terms from common ontologies to name pathway entities significantly reduces the ambiguity that can exist between different pathway representations regarding the named entities, thereby facilitating their integration into a single network for analysis. The use of common ontologies also facilitates the integration of pathways with other kinds of ontology-annotated data.

The use of ontological relations to specify the structure of pathways provides for the direct integration of pathways with ontologies and creation of a unified network that includes the ontology and pathway relations. We anticipate that such a unified network will support the use of the ontology hierarchies to further ease the difficulties of integrating heterogeneous pathway representations.

The ability to incorporate into pathway representations relations asserted in ontologies could reduce the effort of curating pathways, as many of the needed relations are being incorporated into ontologies. For example, the developers of PRO are adding *has_part* relations to PRO, and the CL developers are adding *has_plasma_membrane_part* assertions to CL.

We have encoded the TLR pathways using OWL and are currently developing algorithms that use the relationships encoded in the ontologies for pathway analysis. We have already seen benefits from using OWL for consistency checking to detect curation errors, and we anticipate significant benefit from its application to the detection of inconsistencies in integrated pathway representations. We are also utilizing OWL reasoning to support pathway queries and are currently evaluating the advantages it offers over the keyword querying available through most pathway resources.

The primary barrier we faced in applying this approach to the representation of the TLR pathways was the absence of software. The creation of representational artifacts built from portions of multiple ontologies would be improved by software that allows a user to:

- query a specific set of ontologies, select terms for import, and import specific pieces of information about the term;
- submit term requests to a specific ontology when a needed term cannot be found; and
- assert relations between the imported terms.

The availability of such software would allow this approach to be widely applied.

We see two possible disadvantages to this approach. Curation into the representation framework we describe may take longer or be less intuitive for domain experts than alternative representation frameworks. We believe that any disadvantage in this regard can be addressed through the development of curation software. A second possible disadvantage is a loss of expressivity through the exclusion of domain-specific relations. We believe that the opportunity to directly

integrate portions of ontologies with pathways and compute over the integrated resource provides a benefit that outweighs any possible loss of expressivity. If, however, such a loss of expressivity did present a significant disadvantage, the ontology-based core representations could be enhanced with domain-specific information that may not be accessible to all analysis tools.

Acknowledgments

This work was supported by an NIAID-funded R01 (AI077706) and a Burroughs Wellcome Fund Career Award to LGC.

References

1. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration*. Nat Biotechnol, 2007. **25**(11): p. 1251-5.
2. Smith, B., et al., *Relations in biomedical ontologies*. Genome Biol, 2005. **6**(5): p. R46.
3. Natale, D.A., et al., *Framework for a protein ontology*. BMC Bioinformatics, 2007. **8 Suppl 9**.
4. Meehan, T.F., et al., *Logical development of the cell ontology*. BMC Bioinformatics, 2011. **12**: 6.
5. Finn, R.D., et al., *The Pfam protein families database*. Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
6. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nat Genet, 2000. **25**(1): p. 25-9.
7. Bard, J., S.Y. Rhee, and M. Ashburner, *An ontology for cell types*. Genome Biol, 2005. **6**(2): p. R21.
8. Diehl, A.D., et al., *Hematopoietic cell types: Prototype for a revised cell ontology*. J Biomed Inform, 2010.
9. Rosse, C. and J.L. Mejino, Jr., *A reference ontology for biomedical informatics: the Foundational Model of Anatomy*. J Biomed Inform, 2003. **36**(6): p. 478-500.
10. Masci, A.M., et al., *An improved ontological representation of dendritic cells as a paradigm for all cell types*. BMC Bioinformatics, 2009. **10**: 70.

Bridging Multiple Ontologies: Representation of the Liver Immune Response

Anna Maria Masci¹, Jeffrey Roach², Bernard de Bono³, Pierre Grenon³, Lindsay Cowell⁴

¹Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA

²Research Computing Center, University of North Carolina, Chapel Hill, NC, USA

³European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK

⁴Department of Clinical Sciences, University of Texas Southwestern Medical Center at Dallas, TX, USA

Abstract Chronic liver injury resulting in cirrhosis is the seventh leading cause of disease-related death in the USA. The liver immune response plays a key role in promoting, attenuating, and eventually resolving this severe condition. Although the liver immune response has been extensively investigated by conventional and *omic* approaches, gaps in knowledge still persist, and there is an urgent need for tools to facilitate the integration and analysis of these large, heterogeneous data sets. While ontologies are now accepted as an essential tool for data integration, no currently available ontology includes a representation of immunological reactions in the context of the liver. To address this need, we propose the Liver Immunology Ontology (LIO). LIO is being developed within the Open Biomedical Ontologies (OBO) Foundry framework, importing and linking relevant portions of orthogonal reference ontologies. LIO, is a novel tool for comprehensive analysis of liver immunology data sets, providing a valuable resource for the liver disease research community.

Keywords: Ontology, OBO Foundry, Immunology, Liver diseases, OWL.

1 Introduction

Over the last decades, an increasing amount of evidence has accumulated supporting a pivotal role for the liver immune system in the pathogenesis of severe acute and chronic hepatic disorders, including fibrosis and its final clinical stage, cirrhosis [1]. Critical to uncover the precise mechanisms of this pathogenesis is understanding the unique features of immune response regulation within the liver, in particular through interaction between parenchymal (hepatocytes) and non parenchymal cell types (liver sinusoidal cells, Kupffer cells, hepatic stellate cells, dendritic cells, T cells, natural killer cells, biliar epithelial cells) [2]. A considerable amount of conventional and *omics* findings has been generated from these cell types, but a coherent view of the structural and functional connections between them will require the integration and joint analysis of distinct and heterogeneous data sets.

The annotation of multiple bodies of data using common controlled vocabularies or

‘ontologies’ represents a powerful tool for integrating heterogeneous data sets [3]. Human readability together with machine accessibility make ontologies the perfect tool for organizing, retrieving, and integrating biological data. Although there are well known reference ontologies providing terms relevant to portions of the liver immunology domain, none of them provide a coherent, integrated representation of liver physiology and pathology. Current representations cover only narrow sub domains and are restricted to the assertion of classification and parthood relations within these sub domains. For example, the Foundational Model of Anatomy (FMA) covers normal anatomical entities [4]; the Cell Ontology (CL) covers normal cell types [5-7]; the Gene Ontology (GO) includes a small number of normal developmental and metabolic liver processes [8]; and the Disease Ontology (DO) includes a classification of liver diseases [9]. No currently available ontology includes a representation of immunological reactions in the context of the liver, of the pathological entities relevant to liver disease,

and of the relationships between the various relevant types of entities. To address this gap we are developing the Liver Immunology Ontology (LIO), an application ontology developed within the framework of the Open Biomedical Ontologies (OBO) Foundry [3].

In this paper, we briefly present the approach taken to developing LIO as a product of the selection, combination, and application-specific expansion of relevant communal reference ontologies. We give examples based on the current state of development and discuss briefly our more ambitious aims.

2 Methods

The development of LIO relies heavily on importing terms from existing reference ontologies in order to reduce duplication of effort, better ensure interoperability with other resources, and adhere to the OBO Foundry principles of ontology development best practice [3]. New terms are defined as needed and submitted to the appropriate reference ontology or maintained locally. To provide a coherent, integrated representation of liver physiology and pathology, including of the structural and functional connections between the various relevant entity types, LIO enriches the imported hierarchies by asserting additional relations, in particular between terms imported from distinct reference ontologies.

To date, the development of LIO has focused:

- i) the types of cell found in the hepatic acinus,
- ii) the types of molecule expressed by such cells in the hepatic acinus microenvironment,
- iii) the processes in which these cells and molecules participate, including,
- iv) the outcome of these processes.

To bring together the relevant ontological classes, LIO integrates the high-level classes of five OBO Foundry reference ontologies under the framework of the Basic Formal Ontology.

- The Foundational Model of Anatomy (FMA) (4): *anatomical entities*
- The Cell Ontology (CL) [7, 10]: *cell types*
- The Protein Ontology (PRO) [11]: *proteins*
- Gene Ontology Cellular Component

(GO:CC) [8]: *cellular components*

- Gene Ontology Biological Process Ontology (GO: BP) [8]: *processes*

An extensive review of the primary literature was used to generate a list of all relevant entities. The National Center for Biomedical Ontology Bioportal (<http://bioportal.bioontology.org/>) and the Ontology Lookup Service (<http://www.ebi.ac.uk/ontology-lookup/>) were then used to identify corresponding terms in each of the reference ontologies. When relevant terms were found, they were imported into LIO. When needed terms were not found, specific additions were made to LIO¹.

For each new term, LIO includes a formal definition and a Pub Med identifier supporting the definition. Where appropriate, terms and their definitions were submitted to the relevant reference ontology. The reference ontologies that serve as a source of terms for LIO all use the *is_a* relation as defined in the OBO Foundry Relation Ontology (RO) [12] to form their hierarchy, greatly facilitating the integration of their terms into a single, coherent hierarchy. To capture the structural and functional connections between the various entity types, the integrated LIO hierarchy is enriched with additional relations from the RO and its proposed extension (<http://www.obofoundry.org/ro/>), as described below.

3 Results

Using the approach described above, we are developing an ontology of the immune response induced during liver injury. Immune responses are in general highly context dependent and cannot be described without specifying a well-defined environment. The liver immune response in particular is characterized by local regulation mediated through interaction between parenchymal and non parenchymal cells. Thus, the focus of LIO is inclusion of the structural and functional relationships between these cell types and their components. This is being accomplished, as described above, by

¹ LIO is being encoded using the Web Ontology Language (OWL) (<http://www.w3.org/TR/owl-features/>) and Protégé (<http://protege.stanford.edu/>).

enriching a basic hierarchy, created by importing portions of reference ontologies, with a complex set of intra- and inter-ontological relations. The basic hierarchy of independent continuants is created through importing terms from the FMA, CL, GO CC, and PRO. The hierarchy is enriched to specify the relevant structural features of the hepatic acinus by asserting additional relations as described below. In the example triples that follow, ontology terms are shown in normal font. The source ontology is indicated as a subscript. RO and proposed RO relations are shown in italics.

Some of the anatomical structures that are relevant to the domain of LIO belong to a reference ontology of anatomy. For example, we can find in the FMA terms for the *space of Disse* (FMA ID:63162), *hepatic lamina* (FMA ID:14655) and *endothelium of hepatic sinusoids* (FMA ID:63137). To thoroughly characterize those aspects of the hepatic acinus relevant for describing liver immune responses, additional information is needed. For example, the FMA does not account for the fact that the space of Disse is an anatomical space between the hepatocyte lamina and the endothelium of hepatic sinusoid. Using *adjacent_to* relation from RO, we can capture some aspects of the spatial configuration:

Space of Disse_{FMA} *adjacent_to* Hepatic Lamina_{FMA}

Space of Disse_{FMA} *adjacent_to* Endothelium of Hepatic Sinusoids_{FMA}

LIO requires more than anatomy. In particular, further details of the hepatic acinus microenvironment can be represented by specifying the cellular and molecular components of the environment. This is done by relating anatomical entities (from the FMA) and cellular and molecular entities (from CL and PRO). For example, the space of Disse is characterized by the presence of hepatic stellate cells (cells playing a key role in liver injury, repair and inflammation):

Hepatic Stellate Cell_{CL} *contained_in* Space of Disse_{FMA}

The cell types that participate in and regulate liver immune responses are characterized in LIO by specifying their molecular and cellular component parts. In general, the *has_part* relation is used. Cell-surface molecules, however, are specified using the *has_plasma_membrane_part* relation, which is defined as a sub-relation of the *has_part* relation. Examples include:

Platelet_{CL} *has_part* platelet alpha granule_{GOCC}

Kupffer cell_{CL} *has_plasma_membrane_part* CCR2_{PRO}

Platelet alpha granule_{GOCC} *has_part* VLA-2_{PRO}

In addition to enriching the set of relations between independent continuants in LIO, we enrich the relations between independent continuants and the processes in which they participate. This is critical for capturing the functional relationships at play in a liver immune response. To enable precise characterization of the processes, they are defined as liver-specific processes. They are then related to general terms in GO BP using the *is_a* or *part_of* relation. The processes are then characterized using the *has_participant*, *has_output* and *occurs_in* relations, which are used to specify the independent continuants that are participants in a process, downstream processes that are triggered a process, and the locations in which processes occur, respectively. For example, apoptotic cell clearance is one of the main mechanisms involved in maintaining the homeostasis of the liver immune response. In the liver, Kupffer cell are the cell types involved in this process. This information is captured in LIO using the following set of triples:

Liver apoptotic cell clearance_{LIO} *is_a* apoptotic cell clearance_{GOBP}

Liver apoptotic cell clearance_{LIO} *has_participant* Kupffer cell_{CL}

Liver apoptotic cell clearance_{LIO} *occurs_in* Space of Disse_{FMA}

Kupffer cell cytokine secretion_{GOBP} *has_output* IFN gamma_{PRO}

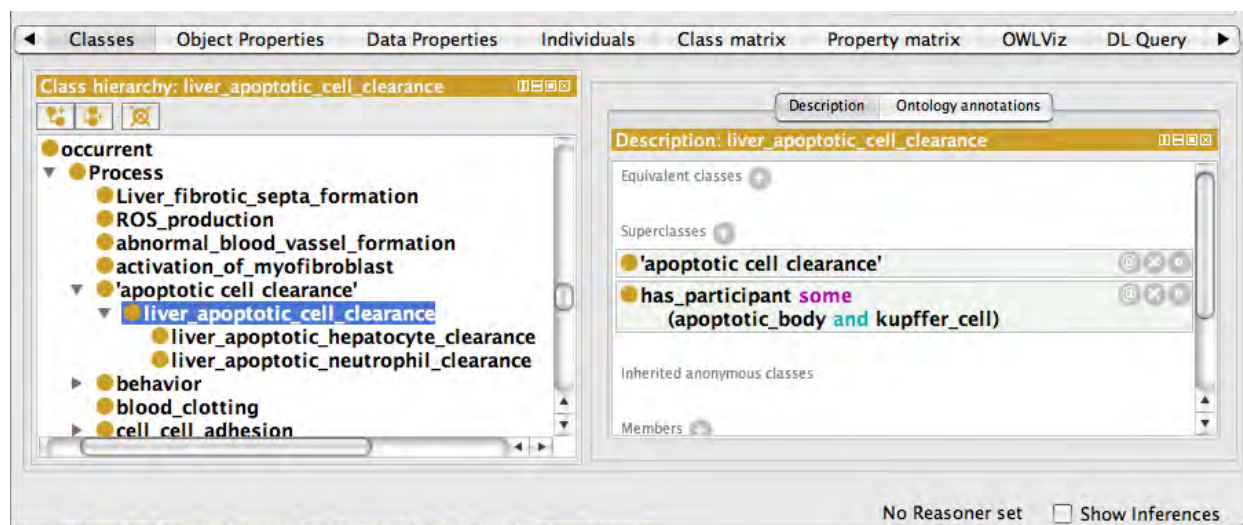


Figure 1. LIO Ontology. A screenshot of the ontology editor Protégé 4 showing the liver apoptotic cell clearance.

The resulting representation creates a bridge of knowledge that spans distinct reference ontologies, ontological types, and multiple levels of granularity.

Assertions in LIO are curated on the basis of the relevant primary literature and supported by associated PMIDs and the evidence code from the Evidence Code Ontology (http://www.obofoundry.org/cgi-bin/detail.cgi?id=evidence_code).

4 Discussion

By representing an immune response in the context of a specific organ, the LIO aims at expanding the resources available for knowledge representation in the domain of liver immunology. Our general approach can be summarized in three steps:

- i) selection of relevant ontology terms from communal reference ontologies to bootstrap our application ontology,
- ii) expansion to fill domain specific gaps after selection,

- iii) combination of ontology terms obtained through selection and expansion to capture a more detailed description of the phenomena and processes that may require crossing multiple domains and levels of granularity.

The examples presented above are relatively simple illustrations of this approach.

They are examples of representation of ground facts. Our aim is to be able to use and refine the sort of knowledge so recorded to define more dynamically the further refinements in the LIO taxonomy. In particular, this is to support the definition of special subclasses of basic classes of liver physiological and immunological processes of LIO in application contexts; for example, kinds of connected pathologies under specific circumstances. Figure 2 illustrates such a use case in which LIO would allow capturing the difference in the physio-pathological responses to different concentrations of the bacterial lipopolysaccharide (LPS) in liver. The result is an explicit and specific linkage between physiological and pathological mechanism.

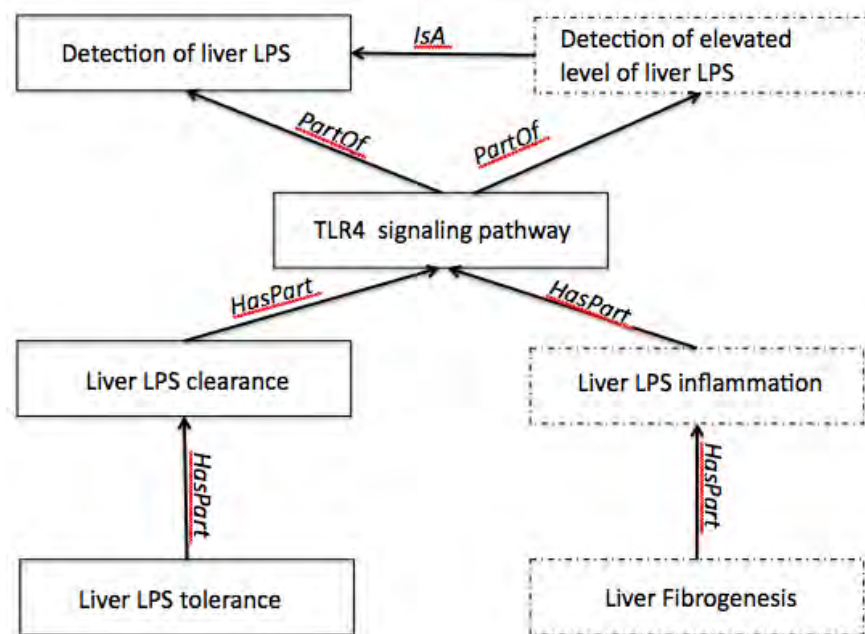


Figure 2. Example of physiological and pathological processes representation in LIO. Shown in the rectangular are processes involved in physiological and pathological conditions (*solid and dotted line, respectively*). The arrows between boxes represent which entities are linked together. In italic are the specific relations used.

We believe that LIO may help illustrate the potential of ontological representation in supporting the scientific understanding of liver immunology. Such a potential, however, can be fully achieved only through the development of an application ontology integrating multiple reference ontologies. On the one hand, such integration allows for a detailed and tailored representation in a specific domain. On the other hand, because such a representation is based on more general reference ontologies, it preserves and benefits from their overall unifying character. By doing so, such a representation also preserves the capacity to be connected to other domains. This is because integration with other domains is facilitated from design at least, as long as the treatments of other application domains are also developed in a similar way.

The resulting benefits for connected research fields, such as the one concerned with liver immunology in the case of LIO, come from the explicit and specific character of the ontology and its capacity to integration from design. This is because bridging multiple ontologies creates a complex network of knowledge able i) to better reflect biological reality and also ii) to maintain a formal

structure required for computational analysis, in general, and, in particular, the interoperability of high throughput technologies on which such analysis is based. Ways of evaluating this hypothesis are dual. First, the correctness of LIO and the adequacy of its coverage of the relevant domain can be evaluated in collaboration with domain experts. Second, the ability of LIO to facilitate data analysis can be evaluated in application driven tests and prototypes.

In the long term, our aim is to use LIO in order to support more comprehensive analysis of existing data than has heretofore been possible, resulting in novel insights, and the generation of new hypotheses. For example, in virtue of its structure, LIO supports an enhanced analysis of microarray data. This is because annotations using LIO would carry, if not no ambiguity at all, at least less ambiguities than context-insensitive annotations using the more general ontology.

Another application of LIO that we wish to explore is to support the biological annotation of relevant mathematical models and associated computer simulations. LIO provides, on the side of reality, the elements of an account of the liver's immune response: the objects involved (cells and molecules), the

environment in which they evolve (liver anatomy), and the relations between these objects and their environment. It provides therefore ontological resources readily conjoined to the mathematical modeling of the liver's immune response, both from the physiological and the pathological perspectives. At stake here is the increased rate of translation of modeling and simulation endeavors into clinical studies, for example, aimed at identifying genes involved in the progression and reversion of liver diseases.

The development of LIO is ongoing and much remains to do in the near future. Our next step will be concerned with the expansion of the representation of the immune response in relation to a variety of triggers such as, for example, infectious agents, alcohol abuse, obesity, autoimmunity and drugs. We believe that furthering the line of development sketched here and extending LIO to more connected domains and resources will provide a valuable resource, meeting the needs of the hepatic disease research community.

Acknowledgments

This work was supported by NIAID R01 AI077706 to LGC (AMM, PI subcontract to Duke University) and a Burroughs Wellcome Fund Career Award to LGC.

References

1. Tacke F, Luedde T, Trautwein C. Inflammatory pathways in liver homeostasis and liver injury. *Clin Rev Allergy Immunol*. 2009 Feb;36(1):4-12.
2. Crispe IN. The liver as a lymphoid organ. *Annu Rev Immunol*. 2009;27:147-63.
3. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007 Nov;25(11):1251-5.
4. Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform*. 2003 Dec;36(6):478-500.
5. Bard J, Rhee SY, Ashburner M. An ontology for cell types. *Genome Biol*. 2005;6(2):R21.
6. Diehl AD, Augustine AD, Blake JA, Cowell LG, Gold ES, Gondre-Lewis TA, et al. Hematopoietic cell types: prototype for a revised cell ontology. *J Biomed Inform*. 2011 Feb;44(1):75-9.
7. Meehan TF, Masci AM, Abdulla A, Cowell LG, Blake JA, Mungall CJ, et al. Logical development of the cell ontology. *BMC Bioinformatics*. 2011;12:6.
8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000 May;25(1):25-9.
9. Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, Zhu LJ, et al. Annotating the human genome with Disease Ontology. *BMC Genomics*. 2009;10 Suppl 1:S6.
10. Masci AM, Arighi CN, Diehl AD, Lieberman AE, Mungall C, Scheuermann RH, et al. An improved ontological representation of dendritic cells as a paradigm for all cell types. *BMC Bioinformatics*. 2009;10:70.
11. Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, et al. The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res*. 2011 Jan;39(Database issue):D539-45.
12. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. *Genome Biol*. 2005;6(5):R46.

Hyperontology for the Biomedical Ontologist: A Sketch and Some Examples

Oliver Kutz¹, Till Mossakowski^{1,2}, Janna Hastings^{3,4}, Alexander Garcia Castro⁵, Aleksandra Sojic⁶

¹Research Center on Spatial Cognition, University of Bremen, Germany

²DFKI GmbH Bremen and University of Bremen, Germany

³Chemoinformatics and Metabolism, European Bioinformatics Institute, Cambridge, UK

⁴Swiss Center for Affective Sciences, University of Geneva, Switzerland

⁵University of Arkansas for Medical Sciences, USA

⁶European School of Molecular Medicine, Milan and University of Milan, Italy

Abstract. The Hyperontology framework has been recently introduced to provide a general methodology for heterogeneous ontology design, i.e. the construction of ontologies that have parts, or modules, written in different formalisms, and which are interlinked in complex ways. We here present a brief outline of this framework, discuss its features and merits, and illustrate its usefulness for the domain of biomedical ontology design by providing and discussing a number of examples.

1 Introduction and Motivation

Ontologies are today being applied in virtually all information-rich application areas, and in particular are of increasing importance in the Life Sciences [16].

While the OWL standard [18] has led to an important unification of notation and semantics, still many diverse formalisms are used for writing ontologies. Some of these, such as RDF, OBO [20] and UML, can be seen more or less as fragments and notational variants of OWL, while others, like F-logic and Common Logic, clearly go beyond the expressiveness of OWL. Moreover, not only the underlying logics are different, but also the modularity and structuring constructs, and the reasoning methods.

Many (domain) ontologies are written in description logics (DLs) such as *SROIQ(D)* (underlying OWL 2 DL) and its fragments. These logics are characterised by having a rather fine-tuned expressivity, exhibiting (still) decidable satisfiability problems, whilst being amenable to highly optimised implementations. However, there are many cases where either weaker DLs are enough – such as sub-Boolean *EL* (an OWL ‘profile’) – and more specialised (and faster) algorithms can be employed, or, contrarily, the expressivity has to be extended beyond the scope of standard DLs.

For example, a weaker DL suffices for the NCI thesaurus (containing about 45,000

concepts) which is intended to become the reference terminology for cancer research [17], but beyond DL expressivity is required for many foundational ontologies, for instance DOLCE [6], BFO¹, or GFO². Note however that these foundational ontologies also come in different versions ranging in expressivity, typically between OWL and first-order or even second-order logic (see Section 3.3 for a discussion of the kinds of problems this entails).

While the web ontology language OWL has evolved and extensions are being constantly developed, its main target application is the Semantic Web and related areas, and it can thus not be expected to be fit for any purpose: there will always be new, typically interdisciplinary application areas for ontologies where the employed (or required) formal languages do not directly fit into the OWL landscape. Heterogeneity (of ontology languages) is thus clearly an important issue. This does not only include cases where the expressivity of OWL is simply exceeded (such as when moving to full first-order logic), but, ultimately, also cases where combinations with or connections to formalism with different semantics have to be covered, such as temporal, spatial, or epistemic logics.

Biomedical ontologies in particular face the problem of heterogeneity, as the information

¹ See <http://www.ifomis.org/bfo/>

² See <http://www.onto-med.de/ontologies/gfo>

that is relevant for such ontologies comprises different data sources such as clinical and experimental data from various epistemic settings. For example, we consider the domain of diseases. A patient's information might include age, family history of disease and social status on the one hand, and on the other hand experimental data for that patient might include metabolic profiles, tumour and genetic markers. Therefore, ontologies of disease need to stretch from an epidemiological, through a traditional clinical representation, to the ontology that includes specific molecular pathways and interactions. In particular, ontologies for complex diseases such as cancer have to deal with spatio-temporal heterogeneity, combinations of qualitative and quantitative data, and missing links between physiological and pathological data [4].

Moreover, in biomedical domains many unknowns still remain, and the questions and theories that drive experimental research also shape the spatial and temporal boundaries of representation. For example, whether mitochondria are classified as organisms or as cellular components depends on the background theory that is accepted, i.e. whether mitochondria are prokaryotes living within eukaryotic cells. Thus, heterogeneity in the ontologies might originate not only from the different formalisms used, but also from heterogeneity within and across specific domains. Therefore, an ontology integration that is intolerant to ontological heterogeneity might not only be unfeasible in practice but also impossible in principle.

We here suggest a heterogeneous framework for the design of biomedical ontologies, based on the theory of hyperontologies as introduced in [12], which suggests solutions to some of these problems of heterogeneity. In particular:

- (i) We briefly sketch the main features of hyperontologies in Sec. 2, including reasoning support based on the tool HETS;
- (ii) We discuss how these features can in general be used within the context of biomedical ontologies in Sec. 3, focusing on 3 aspects in particular, namely (1) borrowing of tools, semantics, and reasoners via logic translations (2) structuring and modularity, and (3) (heterogeneous) ontology refinement.

- (iii) We finally present several examples from the world of biomedical ontology engineering in Sec. 4, illustrating how the structuring techniques for heterogeneous biomedical ontologies can be used in practice; Sec. 5 summarises and discusses future work.

2 A Very Brief Sketch of the Hyperontology Framework

In the presence of several alternative choices of modelling formalisms, it can be a rather difficult task for an ontology designer to choose an appropriate logic and formalism for a specific ontology design beforehand – and failing in making the right choice might lead to the necessity of re-designing large parts of an ontology from scratch, or limit future expandability. Another issue is the mere size of ontologies making the design process potentially quite hard and error prone (at least for humans), which is particularly a problem for ontologies in the Life Sciences. This issue has been partly cured in OWL by the `imports` construct, but still leaves the problem of ‘debugging’ large ontologies as an important issue, see e.g. [10]. Also, simple operations such as the re-use of parts of an ontology in a different ‘context’ whilst *renaming* (parts of) the signature are not possible in the OWL languages, making it difficult to combine ontologies that use the same terms analysed from different modelling perspectives, thereby easily yielding inconsistencies when performing naive ontology combination.

We here propose a solution to the above issues based on the concept of *heterogeneity*: facing the fact that several logics and formalisms are used for designing ontologies, we suggest heterogeneous structuring constructs that allow to combine ontologies in various ways, in a systematic and formally and semantically well-founded fashion. Our approach is based on the theory of institutions (which is a sort of abstract model theory) and formal structuring techniques from algebraic specification theory. Its main features are the following, paraphrasing [12]:³

- The ontology designer can use OBO or OWL

³ For technical detail and extensive discussion we have to refer to [12].

to specify most parts of an ontology, and can use first-order (or even higher-order) logic where needed. Moreover, the overall ontology can be assembled from (and can be split up into) semantically meaningful parts ('modules') that are systematically related by structuring mechanisms. These parts can then be re-used and/or extended in different settings.

- Institution theory provides the framework for formalizing 'logic translations' between different ontology languages, translating the syntax and semantics of different formalisms. These translations allow in particular the 'borrowing' of reasoning and editing tools from one logic to another, when appropriately related.
- Various concepts of 'ontological module' are covered, including simple imports (extensions) and union of theories, as well as conservative and definitional extensions.
- Structuring into modules is made explicit in the ontology and generates so-called proof obligations for conservativity. Proof obligations can also be used to keep track of desired consequences of an ontology, especially during the design process.
- Re-using (parts of) ontologies whilst renaming (parts of) the signature is handled by *symbol maps* and *hiding symbols*: essentially, this allows the internalisation of (strict) alignment mappings.
- The approach allows heterogeneous refinements: it is possible to prove that an ontology O_2 is a refinement of another ontology O_1 , formalised in a different logic. For instance, one can check if a domain ontology is a refinement of (a part of) a foundational one. An interesting by-product of the definition of heterogeneous refinements is that it also provides a rather general definition of heterogeneous sub-ontology and of ontology equivalence.

Tool support for developing heterogeneous ontologies is available via the Heterogeneous

Tool Set HETS, which provides parsing, static analysis and proof management for heterogeneous logical theories. HETS visualises the module structure of complex logical theories, using so-called development graphs. For individual nodes (corresponding to logical theories) in such a graph, the concept hierarchy can be displayed. Moreover, HETS is able to prove intended consequences of theories, prove refinements between theories, or demonstrate their consistency. This is done by integrating several first-order provers and model-finders (SPASS, DARWIN), the higher-order prover (ISABELLE), as well as DL reasoners like PELLET and FACT++.

3 What Hyperontology Can Do for Biomedical Ontologies

3.1 Borrowing Reasoning and Editing Tools via Logic Translation

[14] lays the foundation for a distributed ontology language (DOL), which will allow users to use their own preferred ontology formalism whilst becoming interoperable with other formalisms. At the heart of this approach is a graph of ontology languages and translations. In connection with HETS, this graph enables users to:

- relate ontologies that are written in different formalisms (e.g. prove that the OWL version of DOLCE is logically entailed by the first-order version);
- re-use ontology modules even if they have been formulated in different formalisms;
- re-use ontology tools like theorem provers and module extractors along translations

A detailed discussion of the various translational relationships between (almost) all known ontology languages can be found in [14]. We here concentrate on the languages of specific interest for biomedical ontologies, namely OBO, OWL and its profiles, first-and second-order logic, and F-Logic and Common Logic.

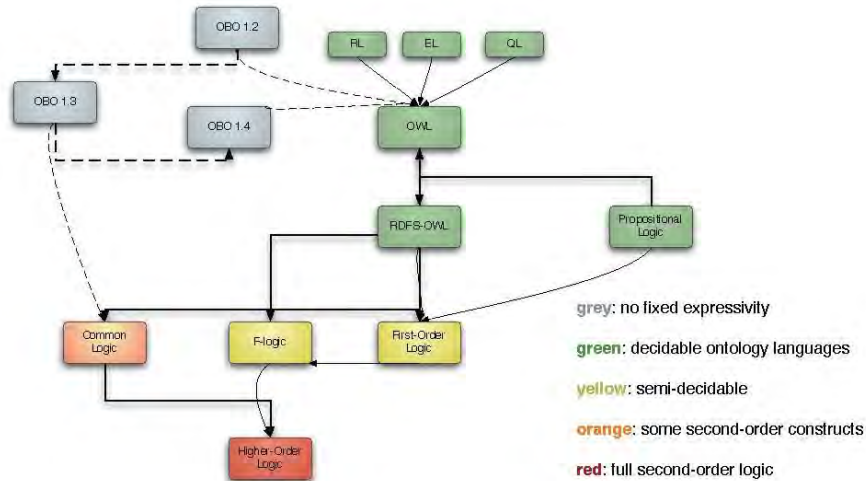


Figure 1. Translations between ontology languages

Fig. 1 illustrates the translational relationships. A ‘regular’ translation between two ontology languages, as marked by a solid arrow, means that the syntax and semantics of one logic can be translated into another. This means that, typically, the former is a fragment of the latter. A standard example would be OWL which, via the standard translation (which is available in HETS), can be considered a fragment of first-order logic. Note that translations concerning different versions of OBO are of different flavours⁴, thus are here marked by dashed lines. The OBO language does not itself come with formal semantics. Beginning with [7], who mapped a fragment of OBO 1.2 to OWL, a semantics for OBO has been assigned by translation. Version 1.3 of OBO, now abandoned, had something similar using Common Logic. The current specification of OBO, version 1.4, gets its semantics entirely via translation to OWL 2. In a sense, thus, the OBO language does not have a fixed logical expressivity, but depends on borrowed model-theoretic semantics from a particular mapping to another ontology language, relative to which corresponding reasoning methods and editing tools can be applied.

Logic translations can in particular be internalised in the ontology languages themselves, in the sense that ontologies can be written in a mix of logical formalisms, where

the translations assign respective semantics by operating in the background. For this to work properly, formal structuring principles are necessary, which we discuss next.

3.2 Structuring and Modularity

The web ontology language OWL as well as OBO can be accommodated within the larger framework of the heterogeneous common algebraic specification language HETCASL. Through this change in perspective, OWL and OBO can benefit from various useful HETCASL features concerning structuring, modularity, and heterogeneity. This tackles a major problem area in ontology engineering: re-use of ontologies and re-combination of ontological modules. We have briefly sketched the main structuring mechanisms already in the last section, namely unions and extensions of ontologies, translations along symbol maps, refinements, conservative extensions, etc.

To be able to write down such heterogeneous ontologies in a concise manner, we propose a structuring language that operates on top of and independently from a chosen ontology language. For instance, we use the notation `logic <logic-name>` to define the logic of the following specifications, which remains intact until that keyword occurs again. Similarly, an ontology (module) can be translated along a logic translation, which is written `<ontology>` with `logic <translation-name>`. The full language, which is also supported by HETS, cannot be given here, but compare [12].

⁴ In particular, the progression between the different versions of OBO are only partial, leaving out some language constructs and adding others.

3.3 Refinements: Relating Domain and Foundational Ontologies

Informally speaking, a (homogeneous) **refinement** of ontology O_1 into another ontology O_2 , both written in the same language, consists of a translation π which translates all of O_1 's axioms in such a way that the translated sentences follow from O_2 . For instance, a Biomedical domain ontology O written in OWL *refines* the OWL version of BFO exactly if O logically implies the translation of BFO's axioms.

But the approach also allows heterogeneous refinements: for instance, it is possible to prove that a first-order version BFO_{FOL} of a foundational ontology, here BFO, is a refinement of an OWL-based version BFO_{OWL} of BFO. Here, it needs to be established, using a first-order theorem prover, that all the translations of BFO_{OWL} 's axioms along the standard translation are logically entailed by BFO_{FOL} . Also, one can check if a domain ontology, written in OWL or OBO, is a refinement of (a part of) a foundational one, written in first-logic. This can be done by first *hiding* a part of the foundational signature, and then establishing a refinement. Note that hiding restricts the vocabulary of an ontology to an "export interface" (which is just a sub-vocabulary), while otherwise keeping the logical properties intact. All these verifications are supported by HETS.

4 Problems and Examples from Biomedical Ontology Design

4.1 Biomedical Imaging

When assessing the mechanical properties of bones, researchers use computational simulations to evaluate stress and strain maps under several boundary and load conditions. Such evaluations involve clinical data, e.g. pathological conditions of the patients, and mechanical properties of the materials to be used. Better models require high quality images, acquisition of which is not an easy task.

There are three main steps when doing computational simulations within the computational biomechanics domain. Pre-processing involves getting the geometrical model of the tissue; medical images obtained by methods

such as Scanning Electron Microscopy (SEM) or Microtomography (μ CT), are the main input for building these models. The images are thresholded, MIMICS can be used for this purpose; the quality of the model is directly related to the level of resolution and number of segmented images. The obtained CAD model offers sagittal, frontal and transversal planes; standard CAD software such as Inventor, Solid Edge, CATIA and Unigraphics are then used to manipulate the model. Finite Elements Methods (FEM) and the post-processing immediately follow; as our aim is to support sharing and reusability of medical images we are only focusing on the pre-processing phase. Details for pre-processing are illustrated in Fig. 2(a).

For describing a medical image it is often necessary to use several ontologies. For instance, Fig. 2(b) (left) illustrates the model for knee joints; hard tissue, e.g. Femur and Tibia bones, and soft tissue, e.g. Tibia and Femur cartilage, need to be identified. The osteoarthritis of the patient requires shaving the cartilage injury as presented in Fig. 2(b) (right). Such self-descriptiveness makes it possible for users to express complex queries such as: '*Retrieve knee joints images with cartilage injury.*'

Facilitating the execution of such a query requires the orchestration of several ontologies, namely: Radiology, Anatomy, Pathology, CAD. Ontologising the description of the DICOM file brings together radiological, spatial, anatomical and pathology related information; since these ontologies are space related, they are not necessarily available as OWL files, and therefore need to be heterogeneously combined using appropriate structuring mechanisms. The degree of segmentation together with other features will determine the suitability of the images. The report of the radiologist and the information contained in the DICOM should support the intelligent retrieval of information as well as the identification and management of anatomical features in the resulting CAD model. To support this with a tool chain is a challenge. Initial experiments in extracting both OWL ABox information and parameters for higher-order specifications from SolidWorks and CATIA CAD models and processing the result with HETS have been made in [11, 5].

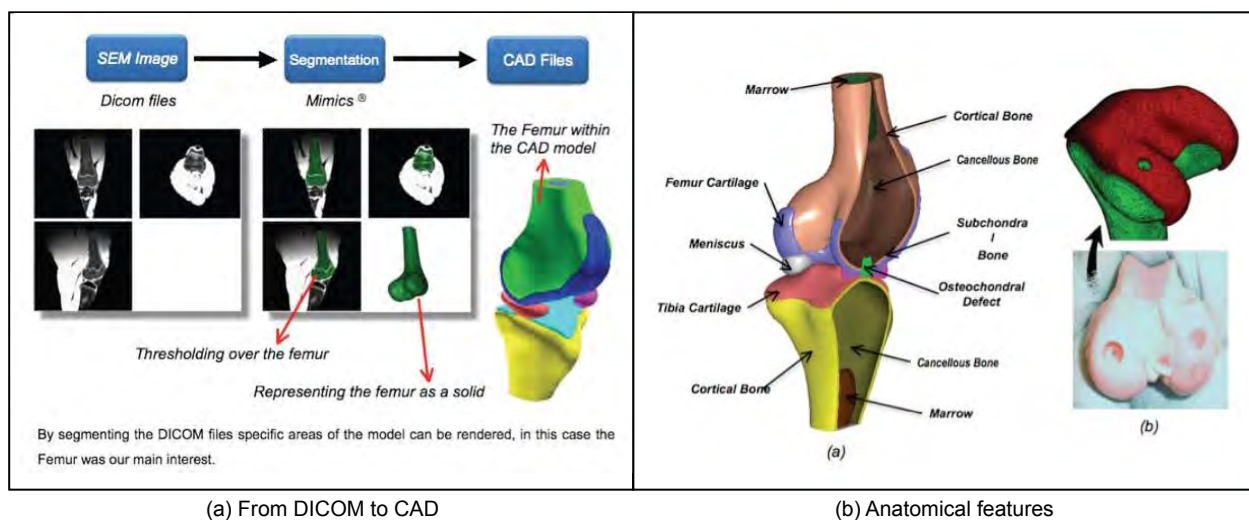


Figure 2. DICOM and Anatomy

4.2 Biochemical Structures

ChEBI (Chemical Entities of Biological Interest) is an ontology of chemical entities consisting of around 25000 entities in the latest release (April 2011) [3]. The core content of the ontology are molecules and ions which are biologically active in some fashion, whether naturally or artificially. The information encoded in the ontology includes a deep structural feature-based hierarchical classification for the chemical entities and a function-based encoding of the actions of the chemicals in biological contexts. For example, *morphine* (CHEBI:17303), an opiate analgesic drug, is included in the ChEBI ontology. It has structural classification, inter alia, in the classes *isoquinolines* and *alkaloids*, and function-based classification in classes *opioid analgesic* and *opioid receptor antagonist*.

Increasingly, OWL semantics is being used for the definition of classes of chemicals based on their shared structural features. Chemical structures are modelled as graphs in which atoms are the vertices and covalent bonds form the edges. For examples of this sort of approach, see [21] and [1]. This allows parts of chemical structures (such as, for example, a *carboxyl group*) to be used to fully define classes (such as, *carboxylic acids*). This is illustrated in Figure 3.

There are several challenges with this sort of approach. The first, well-known, is that chemical entities contain structural cycles. OWL is not suitable for modelling non-tree-like structures, and as a result other formalisms

must be used. Recent work has investigated the use of description graphs and rules to encode these structural features [8], but tool support for description graph-extended OWL ontologies is still poor. In the Hyperontology framework, we would be able to fully describe the structural features in a suitably expressive formalism, and link this to the core OWL ontology, without limiting which tools can be used. More concretely, parts of the modelling can be done in OWL, parts in first-order, and other parts e.g. in a suitable spatial logic, and reasoning is automatically delegated to an appropriate reasoner, possibly employing a logic translation to translate dependent OWL axioms to the first-order formalism.

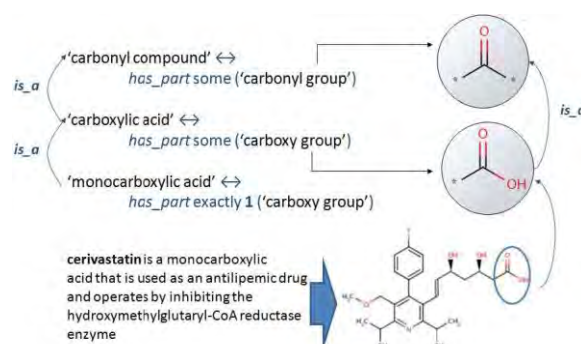


Figure 3. Chemical structures used to define chemical classes

Facing similar challenges, the RNA ontology (RNAO) is an ontology for the structure and function of RNA molecules [9]. RNA molecules consist of chains of nucleotides which can display certain structural motifs or common

patterns. Encoding these structural motifs in the general case requires references to cyclic structures, which can be dealt with in rules or description graphs as for chemicals. Furthermore, the RNAO is provided in a first-order logical formalism implemented in SPASS, and a logically trimmed-down version implemented in OWL. The authors point out that giving a definition for an entirely covalently connected entity such as an RNA molecule based on atoms and bonds would require second-order logic in order to be properly formalised, and for this reason such a formalisation is not provided but the relevant features (transitive closure of the covalent connection relationship) are only approximated in the provided formalism. At present the different versions of the RNAO are not formally interlinked, and each has to be separately maintained and reasoned over. The hyperontology framework allows an elegant solution to this problem, allowing to formally relate different versions of the RNAO, e.g. by heterogeneous refinement (assuming the different versions are logically compatible), and to add second-order constraints on top of weaker formalisations.

4.3 Combining Bio-Ontologies

Since it provides a definition for all biologically interesting chemical entities, ChEBI aims to be sufficient for reuse, allowing use of its axioms as Lego bricks in defining specific molecules and molecular-entity-related biological entities. A quick inspection of ChEBI and a comparison to the related Lipid Ontology (henceforth LO) [1] reveals conceptual relations between the two ontologies. A more detailed review allows us to see the specialization of the axioms in the LO. However, these axioms are not always orthologous to those available in ChEBI, since the LO provides a far more detailed classification for lipids than is the concern of ChEBI. The hyperontology framework allows domain specifications such as the LO to effortlessly re-use parts of core ontologies such as ChEBI and even rename or redefine certain of their entities where needed. Also, more complex relationships between the ontologies' terms can be formalised in a heterogeneous ontology in the style of *Bridge Rules* as they are known from distributed DL or \mathcal{E} -Connections (see [12]).

ChEBI is also used as a reference in biological ontologies. Efforts are underway to explicitly link the Gene Ontology (GO) [19] to ChEBI through the OBO cross-product formalism [15]. Cross-products resemble OWL logical definitions composed in terms of intersection, that is, an 'and' operation. In the first example above, *1,3-dichloro-2-propanol metabolic process* would be formalised in the cross-product style as *metabolic process* **and** *has_participant* **some** *1,3-dichloro-2-propanol*. Several different challenges arise in this ontology alignment process between ChEBI and GO. The first can be characterised as the problem of size explosion: ChEBI and GO are both upwards of 20000 terms with many more relationships, and as a result, reasoning over the combined ontologies can be prohibitively slow. The existing OWL:import mechanism requires the full content of both ontologies be loaded into an application (such as Protégé) in order to work with the cross-ontology definitions. The hyperontology framework allows us to bypass this problem with its built-in support for modularisation, even across ontology languages.

A further difficulty arises because of the common practice of classifying chemicals based on parts of the structure. Under this scheme, ChEBI's nucleotide is classified as a carbohydrate. However, in GO, a biological process such as nucleotide biosynthetic process is not considered to be a subtype of carbohydrate biosynthetic process. The straightforward combination of these two ontologies, ChEBI with its chemical-specific perspective, and GO with its biology-specific perspective, therefore leads to challenges for automated inferences. Present approaches to resolve such difficulties involve lengthy ongoing negotiations between the GO and the ChEBI editors to arrive at mutually satisfactory models that can be shared by both communities. A technological solution which allowed coexistence of the two perspectives without incorrect inferences in either would be better. The hyperontology framework's structuring and linking capabilities offer several roads towards this goal, of course making use of established methods such as statistics-based ontology matching.

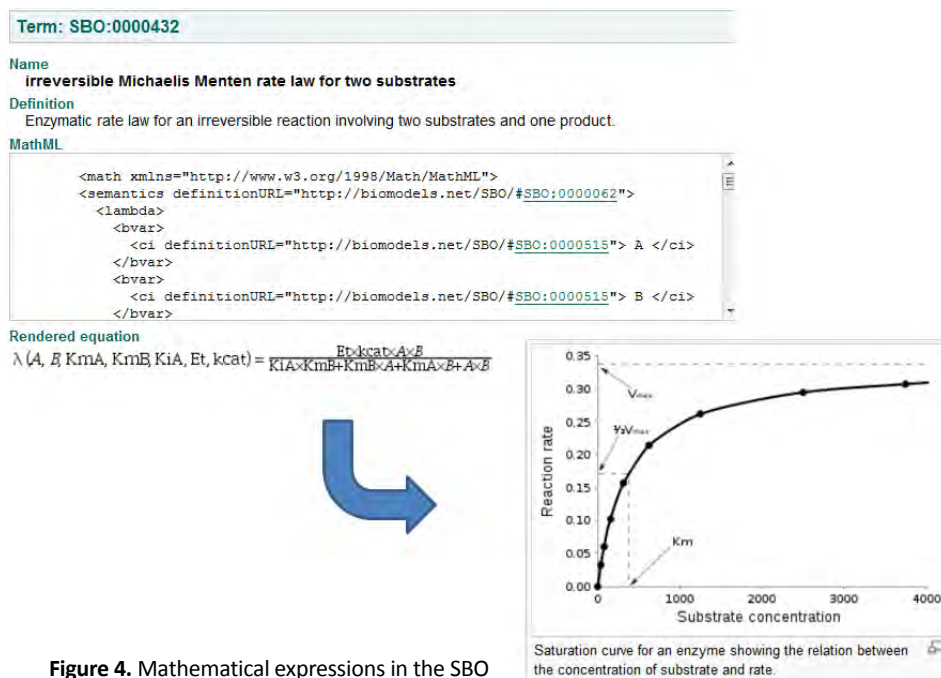


Figure 4. Mathematical expressions in the SBO

Another domain in which the integration of multiple ontologies is mandatory to the creation of successfully interoperable and reusable information is that of Systems Biology. Here, complex mathematical models are used to describe the behaviour of biological systems and to make predictions about their behaviour under different conditions. In order to exchange and unambiguously interpret such models, they need to be annotated with ontologies such as the Systems Biology Ontology (SBO) [13]. SBO contains many different types of entities: material entities such as proteins and small molecules; process participation roles such as inhibitor and stimulant; mathematical laws such as rate laws for biochemical reactions; and types of mathematical model experiments such as discrete and continuous, and many more besides. The hyperontology framework would allow a reformulation of the SBO as composed of modular units sourced from separate domain ontologies, a highly desirable goal. Furthermore, of particular interest in the SBO is that it captures complex mathematical relationships that can exist between biological entities in dynamic conditions. In SBO, these relationships are currently expressed in MathML. It is an open challenge to expose some of the relational information encoded in the SBO mathematical equations to ontological reasoning. This would require interrelating different formalisms,

which is a core feature of hyperontologies.

Describing experimental processes in the biomedical domain also requires a plurality of independent interoperable ontologies. For instance, the Ontology for Biomedical Investigations (OBI) aims to provide a logic conceptual framework for describing biomedical investigations. This task involves interoperability across several ontologies. For example, describing a PCR process involves at least OBI and ChEBI: *buffer*, *reagent* and *phenol* from ChEBI; *thermal_cycler*, *temperature_control_bath* and others from OBI. These classes are usually brought together via either OWL imports of the full ontologies (leading to a size explosion and the accompanying decrease in performance for reasoning) or simply by “slicing” the ontologies and putting together the classes on a need-to-have basis according to the MIREOT methodology [2]. This mechanism is facilitated by tools such as OntoFox [22], which allows users to input terms, fetch selected properties, annotations, and certain classes of related terms from source ontologies and save the results using the RDF/XML serialization of OWL. These hand-selected modules of external ontologies are then brought manually into the target ontology through imports, and the procedure has to be repeated every time the source ontology changes. The hyperontology framework can complement this by a transparent

and automated mechanism to achieve the required interlinked modules, and importantly, extracts modules based on logical principles rather than user steered “slicing”. It remains to be explored how these approaches can be combined and benefit from each other.

5 Outlook

We have sketched the Hyperontology framework and its sophisticated heterogeneous structuring mechanisms, and tried to illustrate their applicability to the domain of Biomedical Ontology by discussing several modelling scenarios in which heterogeneity is a central concern.

Biomedical ontologies, with their complex and heterogeneously interlinked sources of data, conceptual, spatial, and other kinds of knOWledge, probably comprise the most complex application field for ontology engineering today.

In practice, biomedical ontologies often rely on simple subsumption hierarchies (**is_a**), aiming for a generic set of terms and their relationships. Even this, however, requires a strong methodological guidance as most biomedical ontologies are being developed by distributed groups, where teams do not necessarily follow the same classification systems. Such diversity makes a straightforward integration difficult, in particular when moving on to more highly axiomatised ontologies. Biomedical ontologies could evolve and grow in the way they did partly because the use of these controlled vocabularies has mostly been for annotation purposes. In addition, these ontologies are rich in lexical definitions, but not so much in terms of logical axioms. Lexical definitions are surely very important, they make it easier for ontology engineers to elicit knowledge and to manage and understand complex domains. However, a main purpose of formal ontologies is to represent human knowledge so that computers can interpret and reason with it, and not just to facilitate communication across human agents.

When acquiring new knowledge that is beyond the expressivity of is a hierarchies, interoperability across ontologies demands the network of ontologies to rely on a corresponding ‘network of axioms’. Reusing terms from one ontology in a new context cannot rely on simple

‘slicing’ or ‘cutting out’ operations, but has to be based on a logically well-founded method of ‘connecting’ the respective terms. The proposed use of the hyper-ontology framework is intended to enable just such a federated, interoperable solution, focusing on the logical definition of terms and a systematic and structured re-usability of axiomatisations. We are aware that we have only sketched some initial contributions towards this goal; much has to be done to make this really work and get feedback how well it scales in practice.

Acknowledgements

Work on this paper has been supported by the DFG-funded collaborative research centre SFB/TR 8 ‘Spatial Cognition’ and by the German Federal Ministry of Education and Research (Project 01 IW 07002 FormalSafe).

References

1. Baker, C. J. O., Kanagasabai, R., Ang, W. T., *et al.* Towards ontology-driven navigation of the lipid bibliosphere. *BMC Bioinformatics* 9 (2008).
2. Courtot, M., Gibson, F., Lister, A. L., Malone, J., Schober, D., Brinkman, R. R., and Ruttenberg, A. MIREOT: The minimum information to reference an external ontology term. *Applied Ontology* 6 (2011), 23–33.
3. de Matos, P., Alcántara, R., Dekker, *et al.* Chemical Entities of Biological Interest: an update. *Nucl. Acids Res.* 38 (2010), D249–D254.
4. Faratian, D., Clyde, R. G., Crawford, J. W., and Harrison, D. J. Systems pathology – taking molecular pathology into a new dimension. *Nat Rev Clin Oncol* 6, 8 (Aug 2009), 455–64.
5. Franke, M., Klein, P., and Schröder, L. Ontological semantics of standards and plm repositories in the product development phase. In *Global Product Development. Proc. 20th CIRP Design Conference 2010* (2010), A. Bernard, Ed., Springer, pp. 473–484.
6. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. Sweetening Ontologies with dolce. In *Proc. of EKAW 2002* (2002), vol. 2473 of LNCS, Springer, pp. 166–181.
7. Golbreich, C., Horridge, M., Horrocks, I., Motik, B., and Shearer, R. OBO and OWL: Leveraging Semantic Web Technologies for the Life Sciences. In *Proc. of the 6th ISWC* (2007), vol. 4825 of LNCS, Springer, pp. 169–182.
8. Hastings, J., Dumontier, M., Hull, D. *et al.* Representing chemicals using OWL, description

- graphs and rules. In *Proc. of OWLED* (2010).
9. Hoehndorf, R., Batchelor, C., Bittner, T., *et al.* The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structural data. *Applied Ontology* 6, 1 (2011), 53–89.
 10. Kalyanpur, A., Parsia, B., Horridge, M., and Sirin, E. Finding all Justifications of OWL DL Entailments. In *Proc. of ISWC/ASWC* (2007), vol. 4825 of *LNCS*, Springer, pp. 267–280.
 11. Kohlhase, M., Lemburg, J., Schröder, L., and Schulz, E. Formal management of cad/cam processes. pp. 223–238.
 12. Kutz, O., Mossakowski, T., and Lücke, D. Carnap, Goguen, and the Hyperontologies: Logical Pluralism and Heterogeneous Structuring in Ontology Design. *Logica Universalis* 4, 2 (2010), 255–333.
 13. Le Novère, N., Courtot, M., and Laibe, C. Adding semantics in kinetics models of biochemical pathways. In *Proceedings of the 2nd International Symposium on experimental standard conditions of enzyme characterizations* (2007).
 14. Mossakowski, T., and Kutz, O. The Onto-Logical Translation Graph. Tech. rep., University of Bremen, 2011. Submitted.
 15. Mungall, C. J., Bada, M., Berardini, T. Z., Deegan, J., Ireland, A., Harris, M. A., Hill, D. P., and Lomax, J. Cross-Product Extensions of the Gene Ontology. *Journal of biomedical informatics* (Feb. 2010).
 16. Rubin, D. L., Shah, N. H., and Noy, N. F. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics* 9, 1 (2008), 75–90.
 17. Sioutos, N., de Coronado, S., Haber, M. W., Hartel, F. W., Shaiu, W.-L., and Wright, L. W. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics* 40, 1 (2007), 30–43.
 18. Smith, M. K., Welty, C., and McGuinness, D. L. The Web Ontology Language, 2010.
 19. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat. Genet.* 25 (2000), 25–9.
 20. The Gene Ontology Consortium. The OBO language, version 1.2, 2010.
 21. Villanueva-Rosales, N., and Dumontier, M. Describing chemical functional groups in OWL-DL for the classification of chemical compounds. In *Proc. of OWL: Experiences and Directions (OWLED 2007)* (2007).
 22. Xiang, Z., Courtot, M., Brinkman, R. R., Ruttenberg, A., and He, Y. Ontofox: web-based support for ontology reuse. *BMC Res Notes* 3 (2010), 175.

Facilitating Anatomy Ontology Interoperability



ICBO

International Conference on Biomedical Ontology

July 27, 2011
Buffalo, New York, USA

Mouse Anatomy Ontologies and GXD

Terry F. Hayamizu, Martin Ringwald

Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA

Extended Abstract

Anatomy ontologies are essential for the standardized description and integration of many types of biological data pertinent to anatomy, including gene expression, biological process, phenotype and pathology data. From its inception as a community resource for gene expression information for the laboratory mouse, the Gene Expression Database (GXD) has recognized this critical need, actively promoting and contributing to the development of detailed and comprehensive anatomy ontologies for both the developing and postnatal mouse.

The ontology for mouse embryo anatomy was developed by the Edinburgh Mouse Atlas Project (EMAP), with input and feedback from GXD. The EMAP ontology standardizes nomenclature for anatomical structures in the embryonic mouse and provides stage-specific partonomic hierarchies for relationships between developing structures. EMAP ontology development continues as a collaborative effort between EMAP and GXD. The Adult Mouse Anatomy (MA) ontology, developed by GXD, provides standardized anatomy nomenclature for the postnatal mouse, and is organized from both spatial and anatomical system perspectives. The plan is to eventually combine and integrate the ontologies to generate an anatomy ontology covering the entire lifespan of the laboratory mouse.

Both developmental and adult mouse anatomy ontologies have been incorporated into GXD and are currently being used to describe and integrate an extensive range of mouse gene expression data. Currently, GXD includes data for over 50,000 experimental

assays, with greater than 1 million expression results annotated to specific mouse anatomy terms, providing expression information for approximately 12,000 mouse genes. The mouse anatomy ontologies continue to be expanded and refined according to the requirements of accurate and precise data curation by GXD and other resources, as well as in response to input from others in the scientific community.

GXD is an integral part of Mouse Genome Informatics (MGI), which provides integrated access to a wide range of information pertinent to the laboratory mouse. Since many types of data have anatomic attributes, anatomy ontologies are also being used to integrate different kinds of data available for the mouse. For example, Mammalian Phenotype Ontology (MP) development is guided by both MA and EMAP ontologies, by incorporating anatomy term names, confirming term placement within the ontology, and using anatomy terms in logical definitions of MP terms. Furthermore, Gene Ontology (GO) curators use anatomy ontologies as a guide for adding biological process terms that refer to anatomical structures, as well as to indicate specific mouse structures within contextual annotations. Thus, the mouse anatomy ontologies are being used to annotate and integrate gene expression and phenotype data, contributing to an integrated description of biological phenomena in the mouse. Furthermore, we are working on establishing cross-references with anatomy ontologies for other species to support cross-species queries and comparative analysis.

Acknowledgement

GXD is funded by NIH/NICHD grant HD033745.

FUNCARO: A Functional Extension to CARO

David Osumi-Sutherland

Department of Genetics, University of Cambridge, UK

Abstract. Biological structures are commonly classified by functional as well as structural criteria. Here, I propose an upper ontology for functionally defined systems and other functionally defined anatomical structures such as sense organs and glands.

Keywords: function, gene ontology, biological process, anatomy, CARO

1 Introduction

“Structure without function is a corpse; function without structure is a ghost” [1]

Biological structures are commonly classified by functional as well as structural criteria. Some commonly used anatomical terms cannot be defined at all without referring to function: sense organ, gland, endocrine system.

For anatomy ontologies to be interoperable, we need to provide standard ways of classifying anatomical structures according to function. In order to make anatomy ontologies that combine structural and functional classification maintainable, these standards need to be suitable for use in auto-classification by reasoners [2].

The biological process sub-ontology of the gene ontology [3] has a wealth of definitions and classifications for biological processes. With a suitable bridging relation, these can be used to record function.

I propose a draft standard upper ontology, FUNCARO (FUNctional CARO), that combines terms from CARO [4] and the GO biological process ontology to provide a standard framework for functional classification of anatomical structures. I conclude with a discussion of some of the limitations of the approach.

2 Methods

Formal definitions are all given in OWL 2 DL [5], Manchester syntax [6]. **Object properties** are in bold, annotation properties are underlined, *Manchester syntax* itself is italicized.

Given the dominance of OBO 1.2 format in the bio-ontology world, I outline solutions for both OBO and OWL. Throughout this paper I use a nested class expression in OWL of the form “A **has_function** *some (realized_by only P)*” to define the functions of anatomical structures. ro.owl [8] defines a suitable relation for use in OBO ontologies: ‘**has_function_in**’ is defined as an expansion [7] to ‘**has_function** *some (realized_by only Y?)*’.

IDs¹ for ontology terms mentioned:

multicellular organismal process; GO:0032501

detection of stimulus involved in sensory perception; GO:0050906

detection of chemical stimulus involved in sensory perception; GO:0050907

detection of chemical stimulus involved in sensory perception of taste; GO:0050912

secretion; GO:0046903

endocrine hormone secretion; GO:0060986

cortisol secretion; GO:0043400

immune response; GO:0006955

multicellular anatomical structure; CARO:0010000

A draft version of FUNCARO, funcaro.owl, along with required imported files (funcaro_GO_helper_terms.owl and caro_2.owl, which in turn imports terms from PATO_helper.owl) can be found here: <https://arthropod-anatomy-ontology.googlecode.com/svn/trunk/ontologies/trunk/>

¹ IDs are in OBO format, for URI, replace ‘.’ by _ and prepend “<http://purl.obolibrary.org/obo/>”.

3 Results

3.1 Sense Organs

The gene ontology (GO) class ‘detection of stimulus involved in sensory perception’ has 30 subclasses that are fantastically useful for classifying sense organs and sensory neurons. These have been used to define over 1000 classes in the Drosophila anatomy ontology [9]. Unfortunately, CARO does not have a suitable term for organ that can be used as a general genus for classes of sensory organ. We need a term that can refer to small clusters of cells that form most of the sensory organs of arthropods [10] as well as the more complicated sense organs of vertebrates. One possibility is:

label: organ
definition: “A multicellular anatomical structure that is largely delimited by a morphological boundary.”
SubClassOf: ‘multicellular anatomical structure’

But this clearly applies to developing anatomical structures that nobody would refer to as organs. The term “organ” has strong connotations of function. For example, Henderson’s dictionary of biological terms has the definition: “any part or structure of an organism adapted for a special function or functions” [12]. We reflect this functional criterion for class membership by adding a further clause to the definition and adding a functional restriction:

label: organ
definition: “A multicellular anatomical structure that is largely delimited by a morphological boundary and has parts that collectively function in some physiological process.”
SubClassOf: ‘multicellular anatomical structure’
SubClassOf: **has_function** some (realized_by only ‘multicellular organismal process’)

With organ defined, we can now use the GO ‘detection of sensory stimulus’ terms to define and auto-classify 31 sense organ terms. For example:

‘sense organ’ *EquivalentTo*: organ that **has_function** (realized_by only

‘detection of stimulus involved in sensory perception’)

‘chemosensory organ’ *EquivalentTo*: organ that **has_function** (realized_by only ‘detection of chemical stimulus involved in sensory perception’)

‘gustatory organ’ *EquivalentTo*: organ that **has_function** (realized_by only ‘detection of chemical stimulus involved in sensory perception of taste’)

‘detection of chemical stimulus involved in sensory perception of taste’ *SubClassOf* ‘detection of chemical stimulus involved in sensory perception’ *SubClassOf* ‘detection of stimulus involved in sensory perception’

∴ ‘gustatory organ’ *SubClassOf* ‘chemosensory organ’ *SubClassOf* ‘sense organ’

GO also defines sensory perception classes that these ‘detection of stimulus’ classes are part of. If these more general terms are used to define the functions of neurons and neural circuits involved, then we can use reasoning to define perceptual systems.

3.2 Functionally Defined Systems

For functionally defined systems such as the respiratory system, the endocrine system, or the immune system we need to automate population of a partonomy, rather than classification under the system term.

It is useful to define a genus term for functional systems:

label: ‘functional system’
definition: “A material anatomical entity defined by the common function of its component parts. These parts may or may not be connected to form a single structure.”
SubClassOf: ‘material anatomical entity’
SubClassOf: **has_function** some (realized_by only ‘multicellular organismal process’)

Individual functional systems are subclasses of this. For example:

label: endocrine system
EquivalentTo: ‘functional system’ that **has_function** some (realized_by only ‘endocrine hormone secretion’)

A term for components of this system populates the partonomy:

label: endocrine system component
EquivalentTo: 'anatomical structure' *that*
has_function *some (realized_by only*
'endocrine hormone secretion'
SubClassOf: **part_of** *some 'endocrine*
system'

Or, if implementing entirely in OWL, we can replace this with a general class axiom:

'anatomical structure' *that* **has_function**
some (realized_by only 'endocrine
hormone secretion' SubClassOf: **part_of**
some 'endocrine system

With these in place, if we define:

'adrenal gland' *SubClassOf*:
has_function *some (realized_by only*
'cortisol secretion')

'cortisol secretion' *SubClassOf* 'endocrine
hormone secretion'

∴ 'adrenal gland' **part_of** *some 'endocrine*
system'

3.3 Glands

Glands are another type of structure that it is only possible to define functionally. We can define glands as types of organ that function in secretion. For example:

gland *EquivalentTo*: organ *that*
has_function *some (realized_by only*
secretion)

'endocrine gland' *EquivalentTo*: organ *that*
has_function *some (realized_by only*
'endocrine hormone secretion')

'adrenal gland' *SubClassOf*:
has_function *some (realized_by only*
'cortisol secretion')

'cortisol secretion' *SubClassOf* 'endocrine
hormone secretion' *SubClassOf* 'secretion'

∴ 'adrenal gland' *SubClassOf* 'endocrine
gland' *SubClassOf* gland

4 Discussion

4.1 Definition of Organ

The definition of organ here is an improvement on existing purely structural definitions (for example, see the FMA) in that it much better reflects actual usage of the term across species. However, there is still room for improvement, particularly in adding restrictions on the types of boundary that organs can have and in narrowing down the kinds of functions which are required for an anatomical structure to be an organ.

4.2 Potential Problems with Using GO

One of the major challenges to this approach is coordination with the Gene Ontology to make sure that suitable terms are available. In some cases, there is some circularity with Gene Ontology term definitions that we need to resolve. For example, the term 'immune response', which one might expect to be ideal for defining an immune system and its components, references the immune system:

label: immune response
definition: "Any immune system process that functions in the calibrated response of an organism to a potential internal or invasive threat."

The general term 'gland' may also be problematic. GO defines excretion as a subtype of secretion. So by our definition, an excretory organ is a gland. This is certainly not what biologists would expect. One possible way around this is to define glands as sites of both synthesis and secretion, but this is also likely to have exceptions.

4.3 Conclusions

The GO biological process ontology contains a wealth of terms that can be used as functional differentia for defining anatomical classes. Even without additional work, many of these can be used successfully to classify large numbers of anatomical classes. In this paper I have demonstrated the utility if GO terms for functional classification, expanding on an approach that has already been very widely used in the Drosophila anatomy ontology to classify sense organs and sensory neurons [9].

In collecting examples of functional classification in an upper ontology constructed using CARO, FUNCARO provides design patterns for anatomy ontology editors to follow in their work, and so encourages much-needed harmonization of approaches across multiple anatomy ontologies.

In some cases, successful functional classification using GO will require collaboration with GO editors to better define terms. The Gene Ontology editors are very responsive to requests for correction or review of term definitions and their relationships via their tracker [11]. Collaboration with anatomists could also be of benefit to GO – improving and clarifying the terms they have and adding missing terms.

Acknowledgments

I would like to thank: Chris Mungall and Alan Ruttenberg for discussion of how to define function using process terms; the BBSRC (BB/G02247X/1) for funding; and Michael Ashburner for his general encouragement and help.

References

1. Vogel, S. and S.A. Wainwright, 1969. *A Functional Bestiary: Laboratory Studies about Living Systems*. Reading, MA: Addison-Wesley Publishing Co.
2. Rector, A. (2003) Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL, *Proc. K-CAP:ACM 2003*, 121-129.
3. <http://www.geneontology.org/GO.downloads.ontology.shtml>
4. Haendal, M.A., Neuhaus, F., Osumi-Sutherland, D.J., Mabee, P.M., Mejino Jr., J.L.R., Mungall, C.J. and Smith, B. (2007) CARO - The Common Anatomy Reference Ontology: Principles and Practice. In Burger, A., Davidson, D. and Baldock, R.A. (eds), *Anatomy Ontologies for Bioinformatics*. Springer-Verlag.
5. <http://www.w3.org/TR/owl2-primer/>
6. <http://www.w3.org/2007/OWL/wiki/ManchesterSyntax>
7. Mungall, C.J., Ruttenberg, A. and Osumi-Sutherland, D. (2010) Taking shortcuts with OWL using safe macros, *Nature Precedings*.
8. <https://obo-relations.googlecode.com/svn/trunk/src/ontology/ro.owl>
9. http://obo.cvs.sourceforge.net/viewvc/obo/obo/ontology/anatomy/gross_anatomy/animal_gross_anatomy/fly/fly_anatomy_XP.obo
10. Snodgrass, R.E. (1935) *Principles of Insect Morphology*. Cornell University Press.
11. http://sourceforge.net/tracker/?group_id=36855&atid=44076
12. Lawrence, E (Ed) (1995) *Henderson's Dictionary of Biological Terms*, 11th Edition, Longman. Singapore.

The Vertebrate Bridging Ontology (VBO)

Ravensara Travillan¹, James Malone¹, Chao Pang², John Hancock³,
Peter W.H. Holland⁴, Paul Schofield⁵, Helen Parkinson¹

¹EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

²Genomics Coordination Center, Groningen Bioinformatics Center, University of Groningen
and Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands

³MRC Harwell, Harwell, Oxfordshire, UK

⁴Department of Zoology, University of Oxford, UK

⁵Department of Physiology, Development and Neuroscience, University of Cambridge, UK

Abstract. The recent proliferation of ontologies for organizing and modeling anatomical, phenotypic, and genetic information is a welcome development, with a great deal of potential for transforming the way scientists access and use knowledge. Realization of this potential calls for effective ways of integrating and computing on various information sources. In this paper, we introduce the Vertebrate Bridging Ontology (VBO), which permits the transfer of information about homologous anatomical structures between species – a first step towards the integration of species-specific anatomical ontologies. We present the ontology, design patterns, and methodology, and discuss how it can be applied to use-cases to meet the information needs of the scientific user community.

Keywords: anatomy, ontology, vertebrate, evolutionary biology, homology

Availability: <http://sourceforge.net/projects/vbo/>

<http://bioportal.bioontology.org/projects/102>

http://www.ebi.ac.uk/ebiwiki/VBO/index.php/Main_Page

1 Introduction

The problem of integrating diverse single-species anatomy ontologies is well-documented [1]. Comparison of conserved and divergent patterns of gene expression and mutant phenotypes between species has become a powerful approach for investigating gene function and its evolution, particularly as more and more data accumulates from a wide range of species. In order to facilitate a computational approach to cross-species comparisons it is necessary to formalize the description of anatomy in each species, but this then leaves us with the problem of crossing between evolutionarily homologous structures in separate species. Two existing approaches have been attempted: lexical matching and the generation of a “universal” vertebrate anatomy ontology. The former is, for reasons discussed in [1] and below, always going to be intrinsically flawed. The latter has met with some success with the development of the CARO upper level anatomy ontology, and the

Uberon multi-species metazoan anatomy ontology [2, 3]. However neither take full account of the evidence-based inferred evolutionary relationships between anatomical structures in different taxa. In this paper, we introduce the Vertebrate Bridging Ontology (VBO), an evidence – based approach which permits the transfer of information about homologous anatomical structures across species – a first step towards the integration of species-specific anatomical ontologies.

2 Development and Implementation of VBO

The VBO is developed in the Web Ontology Language (OWL) using Protégé 4, in order to provide a common representation compatible with that of the single-species ontologies it is intended to integrate. The OBO (Open Biomedical Ontologies) recommendation of unique namespaces and identifiers has been adhered to in its development.

Use cases collected at a VBO community workshop in June 2010 include key questions the evolutionary-biology and biomedical research communities might wish to address:

1. Gene driven: Compare expression of (a) a named gene or (b) gene family or (c) combination of genes between species in homologous tissues. The queries from this use case will take such forms as: Which anatomical structures are involved in the expression of this {gene | gene family | combination of genes}? Are these structures the same or different in different species? Is expression conserved between species only in homologous structures?
2. Anatomy driven: Compare transcriptomes in a particular homologous anatomical structure between species. The queries from this use-case will take forms such as: For this specific structure, are the same genes or different genes are expressed? What are the differential expression patterns among homologous structures in different species?
3. Compare gene expression similarity and/or difference in particular tissues between species to test a hypothesis of homology. The queries from this use-case will take forms such as: Is Tissue A in Species 1 likely to be homologous to Tissue B in Species 2 as assessed through transcriptome similarity?

Data for these use cases comes from user annotations of model organisms within ongoing human disease mechanism studies, comparative gene expression studies for functional genomics and evolutionary biology, and phenotype/genotype association studies in adult and developing organisms.

Approach. The VBO is based only on anatomical homology – that is, evolutionary relatedness of structures by uninterrupted descent from a common ancestor. The other types of structural similarity in classical comparative anatomy – analogy (similarity of function), and homoplasy (similarity of appearance independent of common descent) – are not part of the VBO scope.

Homology is symmetric, reflexive, and transitive, and thus the homologous nodes for a particular structure form a maximally-

connected graph for the relation *homologous-to*. The combinatorial complexity of the possible axioms linking anatomical entities of even a few species requires a programmatic approach to populating the classes and relationships within the VBO framework. There are two ways to leverage evolutionary anatomical relationships to programmatically populate VBO: a most recent common ancestor (MRCA, “top-down”) approach and a homology chain (“bottom-up”) approach [4], illustrated in Fig. 1.

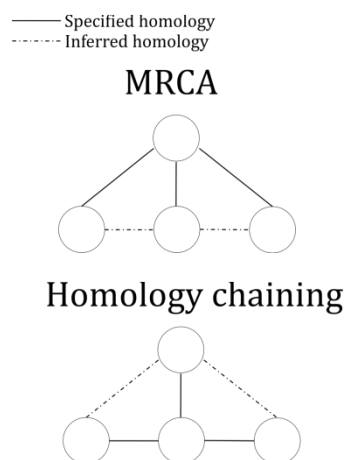


Figure 1. The MRCA approach (*top*) specifies homologies from the MRCA to its descendants, and homologies among the descendants are inferred. The homology chains approach (*bottom*) specifies homologies among the descendants, and requires one explicit connection to the MRCA for that characteristic in order to infer all the other homologies from the descendants to the MRCA.

The two approaches are similar in efficiency, but in principle we favored the MRCA approach as it is more similar to the way biologists reason over evolutionary relationships. In practice, we ended up using a hybrid approach, because the data often were available for one approach but not the other.

Entities. There are two types of entity in VBO: anatomical structures and taxa. An anatomical structure consists of the following data structure (Fig. 2), where the surrounding circles represent annotation properties that link the structure to the homologous structure in other ontologies and taxonomies:

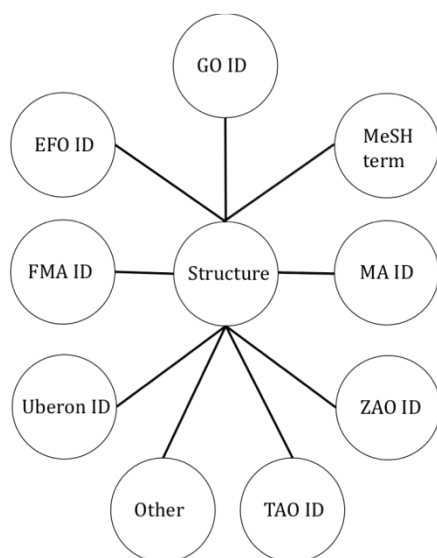


Figure 2. The data structure of an anatomical entity in the VBO (*center*), with annotation properties (*surrounding*).

The corresponding structure(s) in the Experimental Factor Ontology (EFO) [5] is/are linked via the EFO ID, the corresponding structure(s) in the Foundational Model of Anatomy (FMA) [6] are linked via the FMA ID, the corresponding structure(s) in the Teleost Anatomy Ontology (TAO) [7] are linked via the TAO ID, and so forth. The annotation property "Other" represents additional IDs that can be added as the VBO is aligned with additional species anatomy ontologies.

For VBO 1.0, we selected the adult skeletal system for demonstration and proof-of-principle, as it is a relatively straightforward example to model: it tends to be bilaterally symmetrical and highly conserved, with relatively little sexual dimorphism. However, data for other systems became available during the course of the project, so VBO also contains structures outside the adult skeletal system.

Taxon entities can be at any level of phylogenetic ranking, because anatomical structures can be characteristic of any level of ranking. For example, jaws are characteristic of the infraphylum Gnathostomata, while hair, sweat (eccrine) glands, and mammary glands are characteristics of the class Mammalia, and hypertrophied manus digits supporting wings are characteristic of the order Chiroptera. While the scope of the VBO is vertebrate structures, many structures that are characteristic of vertebrates actually originate further back in evolutionary history, so a rigorous

modeling of the VBO requires the ability to model structures as differentia at the appropriate taxon ranking. The current VBO phylogeny is consistent with the NCBI taxonomy for vertebrates.

Taxon entities in VBO have the following data structure (Fig. 3):

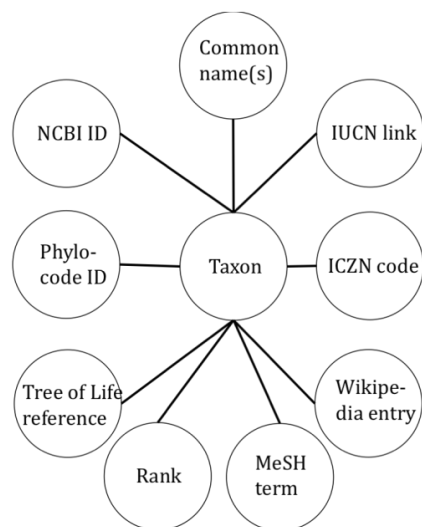


Figure 3. The data structure of a taxon entity in the VBO (*center*), with annotation properties (*surrounding*).

A compound entity represents a structure in a species, as represented in Fig. 4.

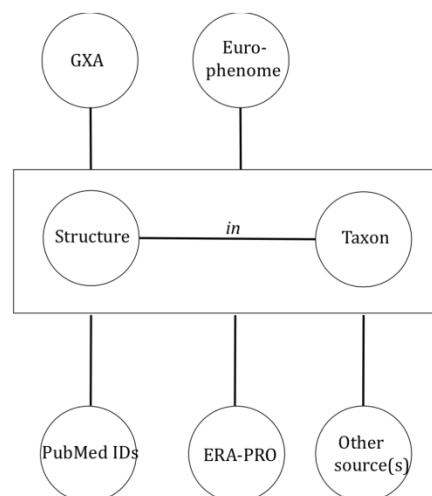


Figure 4. The data structure of a compound entity, representing a structure *in-a* taxon, and the annotation properties that document the evidence of existence of that compound entity: PubMed, ERA-PRO, Gene Expression Atlas (GXA), Europhenome, and so forth.

Compound entities also have annotation properties representing the source of the assertion that

$$\forall \text{Structure-in-Taxon} \rightarrow \{\exists (x \in \text{Taxon}) : (1) \\ x \text{ has-part Structure}\}$$

The following relationships operate on compound entities:

Relationships. These relationships in the VBO describe homology relationships among compound entities.

1. Homologous-to. The relationship homologous-to describes a 1:1 (injective) and onto (surjective) (thus, bijective) structural similarity based on evolutionary relationship between a structure in one species and a structure in a second species, as in Fig. 5.

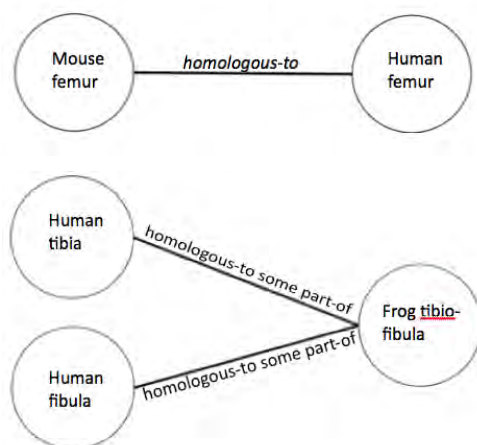


Figure 5. Mouse femur *homologous-to* human femur (top), human tibia *homologous-to* some part of frog tibiofibula and human fibula *homologous-to* some part of frog tibiofibula. (bottom).

While not definitively ruling out a genetic event that occurred after the species' separation from the MRCA, a 1:1 and onto mapping tends to be indicative of evolutionary conservation. When the mapping by term name or structure is not itself 1:1 and onto with a homologous structure (which can indicate an evolutionary event), there may be a 1:1 and onto mapping from a structure in one species to some part of the homologous structure in the second species.

2. Not-homologous-to. The need to explicitly encode a negative relation in VBO is a consequence of the combination of open-world reasoning and the history of comparative anatomy. The *not-homologous-to* relationship can be one-to-many.

The naming of structures in one species, based on analogy ("wing" in insect, pterosaur, bird, and bat) or homoplasy (panda's "thumb") to a non-related structure in a different species, muddies the waters tremendously for determining homology based on lexical matching. Haendel *et al* (accessed 10 April 2011) have remarked upon the case of the frontal bone in the zebrafish being homologous to the prefrontal bone, and not the frontal bone, in humans. The problem is magnified tremendously by the use of important vertebrate skeletal terms to refer to segments in insects, and that is in turn magnified by the importance of those insects, such as *Drosophila*, in the comparative medical research community. Table 1 presents an illustration of the problem for some representative skeletal structures.

Structure	Invertebrate taxa	Refers to	Vertebrate taxa	Refers to
acetabulum	parasitic worms (trematodes), leeches	the sucker (feeding)	tetrapods (4-limbed vertebrates)	concave pelvic surface meeting femur at hip joint
femur	insects	leg segment	tetrapods (4-limbed vertebrates)	long bone in leg
trochanter	insects	leg segment	tetrapods (4-limbed vertebrates)	part of thigh bone
coxa	insects	leg segment	tetrapods (4-limbed vertebrates)	hip (either joint or anatomical region)
tibia	insects	leg segment	tetrapods (4-limbed vertebrates)	long bone in leg

Table 1. Representative identically-named vertebrate and invertebrate non-homologous structures in PubMed.

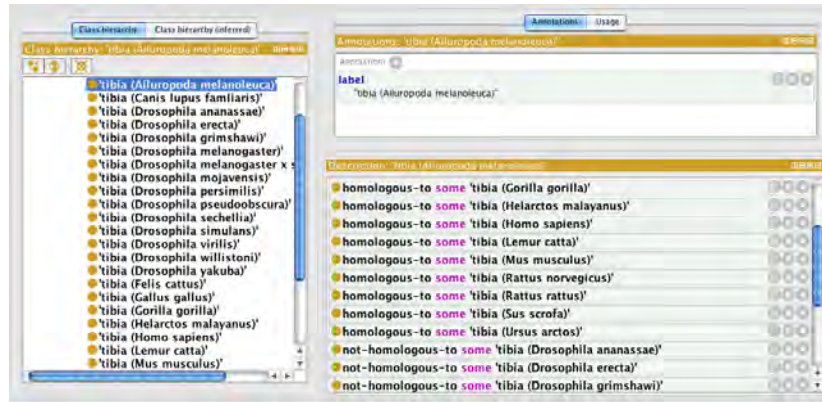


Figure 6. Sample entities and relations in the VBO.

The open-world assumption means that any lexical-matching tool used to populate VBO or any other homology-based ontology will create a high number of false positives based on lexical matches such as these, since – under that assumption – there could, in future, be insect structures that are homologous to their vertebrate homonyms. This possibility, permitted under the open-world assumption, actually violates a biological constraint on homology. To prevent those false positives, to provide metaknowledge for future data mining tools, to mitigate human error in creating axioms containing NOT and a vast number of disjoints in Protégé, and to make reasoning more tractable, we have explicitly encoded the *not-homologous-to* relationship, along with any necessary invertebrate species, in the VBO in order to definitively rule out that possibility. Although it is not an ideal solution, it is a workable compromise, given the state of the art and the scope of the problem. We do not represent a phylogeny of invertebrates, nor do we make any statements about the relationships among *not-homologous-to* relationships, as those are clearly out of scope, so *not-homologous-to* forms a simply-connected graph, and not a maximally-connected one.

Entities and relationships as described above provide the content of VBO; Fig. 6 shows representative entries for a vertebrate tibia, and its relationships to other vertebrate and invertebrate tibiae.

VBO was initially populated by a combination of manual and automated approaches. Annotations from the Gene Expression Atlas [7, 8], ERA-PRO [9], Europhenome [10, 11], and Phenoscope [12] databases provided anatomical structures and

species for the ontology. Additionally, Uberon and FMA provided structures for VBO. These structures and species were manually added to the OWL file in Protégé. For VBO 1.0, inclusion of a taxon or structure class in one of the above databases or ontologies was considered sufficient evidence of existence to include it in the ontology. The use of these sources also uncovered some major discrepancies between how major ontologies, such as FMA and Uberon, represent anatomical classes versus the way the terms corresponding to those classes are used in real-world contexts [1]. Those considerations influenced how we developed composition of compound entities, for example, and will continue to inform future versions of VBO.

Some preliminary data-mining of PubMed abstracts was carried out to populate VBO. Python scripts which searched PubMed iteratively through a list of structures from FMA and Uberon were used to collect abstracts of articles that contained musculoskeletal terms with references to non-human vertebrate species. Reference to a structure in a species in an abstract was considered evidence of a compound entity (Equation [1]), and the compound entity was evaluated for homology to that structure in humans or another species. This evaluation was carried out on the basis of available evidence – reference material, journal articles, and so forth. The provenance of the evidence was recorded as well. This direct connection to evidence for homology statements is a unique strength of VBO.

When sufficient evidence established the homology between the compound entities, the triple

<Compound-entity-1>*relationship*<Compound-entity-2>

was recorded as a “pairwise mapping” in a spreadsheet. A set of Java tools was developed to transform the spreadsheet's pairwise mappings into classes and relationships in Protégé, and to create the relationships among the nodes of the maximally-connected graph. These generated relationships are marked evidentially as inferred from homology.

A beta version of VBO has been successfully integrated into the EFO to support cross-species comparisons of orthologous genes in homologous tissues through the Gene Expression Atlas interface.

3 Future work

We plan to continue integrating VBO into the Gene Expression Atlas via EFO, and improving the functionality and the interface. We will add more sophisticated analysis of evidence that can work with the Phenoscope taxonomy of evidence model for easier integration and sharing of data. More complex systems which present more complicated modeling challenges, and incorporating developmental structures as well as adult structures are also areas into which we plan to extend VBO.

Acknowledgements

We thank the members of the VBO Scientific Advisory Board, Jonathan Bard, Claudio Stern, Martin Ringwald, and Monte Westerfield, who guided and supported this project. In addition, we thank Hilmar Lapp, and the participants in our community workshops, who provided valuable feedback and suggestions.

Funding: Biotechnology and Biological Sciences Research Council (grant #BB/G022755/1), and European Molecular Biological Laboratory core funding.

References

1. Travillian RS, Adamusiak T, Burdett T, Gruenberger M, Hancock J, Mallon AM, Malone J, Schofield P, Parkinson H. Anatomy Ontologies and Potential Users: Bridging the Gap. Biomed Semantics, forthcoming.
2. Haendel M, Gkoutos G, Lewis S, Mungall C. Uberon: towards a comprehensive multi-species anatomy ontology. Available from Nature Preceedings (2009).
3. Haendel MA, Neuhaus F, Osumi-Sutherland D, et al. CARO - The Common Anatomy Reference Ontology. In: *Anatomy Ontologies for Bioinformatics, Principles and Practice* Albert Burger, Duncan Davidson and Richard Baldock (Eds.), 2007.
4. Osumi-Sutherland D. personal communication, 2010.
5. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010 Apr 15;26(8):1112-8.
6. Rosse C, Mejino JL. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003 Dec;36(6):478-500.
7. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. Gene Expression Atlas at the European bioinformatics institute. Nucleic Acids Res. 2010 Jan;38 (Database issue):D690-8.
8. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A. ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009 Jan;37(Database issue):D868-72.
9. Birschwilks M, Gruenberger M, Adelman C, Tapio S, Gerber G, Schofield PN, Grosche B. The European radiobiological archives: online access to data from radiobiological experiments. Radiat Res. 2011 Apr;175(4):526-31.
10. Morgan H, Beck T, Blake A, Gates H, Adams N, Debouzy G, Leblanc S, Lengger C, Maier H, Melvin D, Meziane H, Richardson D, Wells S, White J, Wood J; EUMODIC Consortium, de Angelis MH, Brown SD, Hancock JM, Mallon AM. EuroPhenome: a repository for high-throughput mouse phenotyping data. Nucleic Acids Res. 2010 Jan;38(Database issue):D577-85.
11. Mallon AM, Blake A, Hancock JM. EuroPhenome and EMPReSS: online mouse phenotyping resource. Nucleic Acids Res. 2008 Jan;36(Database issue):D715-8.
12. Dahdul WM, Lundberg JG, Midford PE, Balhoff JP, Lapp H, Vision TJ, Haendel MA, Westerfield M, Mabee PM. The teleost anatomy ontology: anatomical representation for the genomics age. Syst Biol. 2010 Jul;59(4):369-83.

Multi-Species Anatomy Ontology Development Requires a Pluralist Approach to Label-Class Mapping

István Mikó, Matthew J. Yoder, Matthew A. Bertone,
Katja C. Seltmann, Andrew R. Deans

North Carolina State University, Department of Entomology, Raleigh, NC, USA

Abstract. Scientists use labels to reference complex anatomical entities that could, in practice, be represented by spelling out their circumscription in excruciating detail. The referential approach, however, potentially limits the utility of the work by forcing the consumer to interpret the underlying meaning sought by the author. These problems of interpretation are exacerbated when the user communities grow to include experts with disparate training and conventions, *e.g.*, the typical community of multispecies anatomy ontology consumers. Anatomy ontologies are poised to address these problems by providing logically parsable definitions for classes and by reconciling the labels and their contexts. Here we discuss mechanisms used by the Hymenoptera Anatomy Ontology project to solicit broad buy-in and promote harmony amongst a very broad array of users.

Keywords: biomedical ontologies, semantic conflicts, preferred terms, Hymenoptera Anatomy Ontology

1 Representing Common Classes, Rather than Forcing Common Vocabularies

A strong argument for the development of an ontology is that it can stand as a controlled vocabulary that unifies within and cross-domain knowledge [1, 3]. There are obvious benefits to a 1:1 relationship between labels and classes [2]: improved readability and navigability of ontology class hierarchies and the facilitation of alignment and integration. An emphasis on 1:1 label-class relationships, however, inhibits broad buy-in.

Any mechanism that promotes the adoption of ontologies by domain experts is valuable [4]. We propose that conflicts arising from debates focused on “preferred terms” or absolute disambiguation are more likely to occur within multi-species ontologies, in which the spectrum of expertise across domain experts is particularly broad, and in which synonymy and homonymy are more prevalent due to isolation of research in particular sub-domains. In a practical context, ignorance of ontologies by domain experts might easily turn to aversion if they are ordered to follow a

terminology that is represented in a difficult and cutting edge construct (ontology). Eliminating or minimizing this conflict (*i.e.* the emphasis of the importance of a 1:1 relationship) should result in an increased acceptance of the ontology.

The requirement for unique labels also introduces problems across ontologies, which results in the proposal of *ad hoc* solutions, like appending organismal or other context-based strings in front of labels (see discussion on the OBO-listserv from 2010 [5]). Application-level mechanisms exist already for the disambiguation of concepts, *e.g.*, meaningless identifiers, and so debates about unique labels ultimately displace effort that could otherwise be spent on class clarification. Aspiring for ontological exactness using referents that can be trivially misinterpreted will ultimately fail. The term “process”, for example, will never have a single meaning, either functional or anatomical. However, a meaningless, globally unique identifier – ID:012345 – stands a chance of stable mapping. Together these arguments emphasize the importance of a shared set of common classes rather than a common vocabulary.

2 Facilitating Adoption of Multi-Species Anatomical Ontologies – A Pluralist Approach and Currency of Adoption

A pluralist approach to ontology development is one that de-emphasizes the importance the 1:1 relationship between label and classes. The Hymenoptera Anatomy Ontology (HAO) embraces and facilitates this approach via the “sensu” model [6]. A “sensu” is simply a combination of class, label, and reference, which points to an author and the context. A domain-user mechanism for interacting with an ontology employing sensus is, minimally, a text “URI table” with three columns that can be included in a “materials and methods” section of a published work. This table links the annotator’s or author’s text to the ontological meaning intended. The first column is required and contains a label for the annotation or publication. The second column is optional and contains the URI of the class being referenced if the author has found a match in the ontology. In the case of the HAO we have provided a simple search interface that guides the user towards the definition being sought, and, when found, provides the necessary URI. The third column, required only when the URI is empty, contains a definition, preferably presented as a genus-differentia and formatted according to the ontology’s guidelines (though this is not necessary). This definition represents a new class or a proposed correction if the second column is also populated. This type of information is typically already provided within publications authored by multi-species domain experts, like taxonomists or phylogeneticists, the only necessary addition being the inclusion of a URI. A discussion of the table can be included, which would ideally articulate arguments for why definitions were altered or why a label was used in a novel combination.

The URI table acts as currency for interaction between domain experts and ontology curators. It is easily interpreted by the ontology curators, who better understand the logical constraints of the ontology. By facilitating interaction, it further promotes interaction with the ontology, which will ultimately shape the value of the ontology,

and it guides people onto a path towards the ultimate adoption of the central principles for ontology building. Furthermore, when formally included, a URI table acts as the basis for various metrics by which a “preferred” term could be computed, should it be deemed necessary (e.g. number of usages, number of conflicting usages, temporal distribution of use).

A pluralist approach has various inherent advantages. It frees domain expert curators to focus the more important aspects of the ontology – its class definitions and relations – and to act as supervisors rather than dictators. This approach provides domain experts the freedom to use their own labels and to explicitly accept or reject individual classes (*i.e.* logical definitions) of a given ontology without concern as to the label representing that class. In this framework authors are free to propose new labels with confidence, understanding that others will be able interpret their work based on definitions, they are also given full credit for their additions to the ontology (via the sensu).

At its core, a pluralist approach emphasizes the distinction between the referencing mechanisms (IDs, labels), and the underlying meaning of an ontology (its class definitions and relations). Ontologies with definitions that are inconsistently written, poorly conceived, or at worst non-existent, while referenceable, are destined to confuse, or in a worse case scenario, positively mislead. *Domain experts are the primary source of meaningful input* to expert bodies of anatomical data, as such we must minimize the hurdles they must leap to engage an ontology.

We propose that the historical importance of a controlled vocabulary be downplayed, recognizing that there are many application and publication level mechanisms for referencing ontologies that do not require a 1:1 bound of label to class.

References

1. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology by Noy, N.F. and McGuinness, D.L. SMI technical report SMI-2001-0880, Stanford University (2001)

2. Schober, D., Smith, B., Lewis, S.E., Kusnierczy, W., Lomax, J., Mungall, C., Taylor, C.F., Rocca-Serra, P., Sansone, S.A.: Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics*, 10, 125 doi:10.1186/1471-2105-10-125 (2009)
3. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25, 1251–5. doi:10.1038/nbt1346 (2007)
4. Swartout, W.R., Patil, P., Knight, K., Russ, T.: Toward Distributed Use of Large-Scale Ontologies. In: *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop*. Banff, Canada (1996)
5. Multiple authors.
http://sourceforge.net/mailarchive/message.php?msg_id=26794461. Obo-discuss listserv at <https://lists.sourceforge.net/lists/listinfo/obo-discuss> (2010)
6. Yoder, M.J., Mikó, I., Seltmann, K.C., Bertone, M.A., Deans, A.R.: A Gross Anatomy Ontology for Hymenoptera. *PLoS ONE* 5 (12), e15991. doi:10.1371/journal.pone.0015991 (2010)

CARO 2.0

David Osumi-Sutherland

Department of Genetics, University of Cambridge, UK

Keywords: CARO, anatomy, ontology, OWL, PATO

The Common Anatomy Reference Ontology (CARO) [1] has been used extensively as an upper ontology for many anatomy ontologies (Teleostei, Hymenoptera, zebrafish, Drosophila, plants, even Dictyostelium). However, its application has not been very consistent.

There are a number of factors contributing to this inconsistency. Firstly, many of CARO's definitions are quite opaque. In particular, many use specialist terms (e.g. multi-tissue aggregate) that are not defined elsewhere in CARO or in references associated with the definitions. Secondly, it contains no equivalent class definitions or declarations of disjointness, although the original paper [1] makes clear that CARO was intended to be pairwise disjoint. Consequently, CARO cannot be used with a reasoner to aid ontology building via auto-classification and consistency checking. Thirdly, almost all users of CARO have re-implemented its terms in their ontology, rather than use an import system, leading to deviations from CARO in local implementations. Recently there has been increased interest in the development of multi-species anatomy ontologies, including efforts to build or refine anatomy ontologies for arthropods and vertebrates and plants. This work has highlighted many of the problems with CARO and its inconsistent application and made fixing them an urgent priority for the groups involved.

I will present details of a draft revised CARO [2], developed in OWL, that aims to correct these issues. In the new draft:

(a) Definitions have been simplified;

- (b) Equivalent class definitions (intersections in OBO) have been used wherever possible, for example 'material anatomical entity' is defined as *EquivalentTo*: ('anatomical entity' *that has_quality some* PATO:mass);
- (c) Disjointness has been widely declared, for example: 'material anatomical entity' *DisjointWith* 'immaterial anatomical entity';
- (d) Useful terms missing from CARO, such as 'multicellular anatomical structure', have been added and defined;
- (e) Some use has been made of the increased expressiveness of OWL over OBO, for example: 'multicellular anatomical structure' is defined as *EquivalentTo*: ('anatomical structure' *that (has_component min 2 cell)*), where **has_component** is a non-transitive *SubPropertyOf* **has_part**.

References

1. Haendel, M.A., Neuhaus, F., Osumi-Sutherland, D.J., Mabee, P.M., Mejino Jr., J.L.R., Mungall, C.J. and Smith, B. (2007) CARO – The Common Anatomy Reference Ontology: Principles and Practice. In Burger, A., Davidson, D. and Baldock, R.A. (eds), *Anatomy Ontologies for Bioinformatics*. Springer-Verlag.
2. caro_2.owl is available from:
<http://tinyurl.com/6a595gj>
(Loading requires a file of PATO terms: <http://tinyurl.com/6dq6f22>). Instructions for Protégé setup for viewing can be found here: <http://tinyurl.com/69sxt3r>.

Integrating Anatomy and Phenotype Ontologies with Taxonomic Hierarchies

James P. Balhoff^{1,2}, Peter E. Midford¹, Hilmar Lapp¹

¹National Evolutionary Synthesis Center, Durham, NC, USA

²University of North Carolina, Chapel Hill, NC, USA

Cross-species anatomy ontologies and shared organismal taxonomies provide powerful tools for integrating comparative observational data with other knowledge domains. In current practice, organismal taxonomies are typically represented as subclass hierarchies in which taxa are modeled as ontological classes and individual organisms as instances of those classes. However, in contrast to the typological nature of ontological classes, evolutionary data regularly show polymorphism, reversals, and novelty. We have found that modeling taxonomic groups as individuals instead of as classes is more consistent with evolutionary biological data and knowledge, and also has desirable practical consequences when linking observational data such as phenotypes to taxa within the OWL DL framework.

The way taxonomies group organisms into successively lower and lower hierarchical levels is informed by the characteristics that the members of these groups share, and those that distinguish them from each other. It is thus tempting to view a taxonomic hierarchy as a genus-differentia system in which classes can be logically distinguished from one another by the essential features that their members (i.e., instances) must possess. For example, “Tetrapoda are vertebrates that have four limbs”, or, “Mammalia are tetrapods that have hair and mammary glands”. However, the features shared by organisms grouped together in this way are the result of their presumed common evolutionary descent, and it is foremost this shared evolutionary history that organismal taxonomies aim to reflect. Hence, the distinguishing features of a group are assumed to have originated in their common ancestor, but may subsequently among the lineage of descendants become modified, lost, or polymorphic due to the continuing evolutionary process. As David Hull writes, “Organisms belong in a particular species because they are

part of that genealogical nexus, not because they possess any essential traits” [1].

As a consequence, applying a naive typological approach to modeling evolutionary data in a Description Logic (DL) framework such as OWL DL can result in unintended inferences and logical inconsistencies. For example, it is cumbersome to represent polymorphic character states, and expressing evolutionary reversals or loss of phenotypes using cardinality or negation (“has_part only (not scale)”) can result in unsatisfiable classes. If instead we model taxa as individuals, and use object properties rather than subsumption to reflect the taxonomic hierarchy, we can control the propagation of property assertions up and down the taxonomic hierarchy.

For the purposes of demonstrating the advantages of this approach in several examples, we use the transitive object properties *subclade_of* (and its inverse *contains_clade*) to establish the hierarchy of taxa, which are all instances of type *Clade*, and the object property *member_of* (and its inverse *has_member*) to relate individual organisms to a taxon. We focus our examples on anatomical phenotypic traits, which we express as (anatomical) entities that bear some quality, and associate to organisms using the *has_part* relationship. Using these conventions, we can capture the observation of a bifurcated dorsal fin in the channel catfish *Ictalurus punctatus* in the following way:

```
Individual: Ictalurus_punctatus
Types: Clade, has_member some
      (has_part some (caudal_fin and
      (bearer_of some bifurcated)))
Facts: has_rank species,
      subclade_of Ictalurus
```

By defining a property chain *contains_clade o has_member → has_member*, we can infer that the order Siluriformes, which *contains_clade* *Ictalurus* (and thus *Ictalurus punctatus*), has

some member organism with a bifurcated caudal fin. Additionally the set of all such orders can easily be expressed as ‘(has_rank value order) and (has_member some (has_part some (caudal_fin and (bearer_of some bifurcated))))’. Thus, querying for higher taxa based on traits identified in sub-taxa, or in other words up-propagating trait observations to higher taxa, is straightforward in this approach, but difficult to accomplish if taxa are modeled as classes.

In addition to direct observations of traits, a common result of comparative analyses is the assertion of certain shared traits as ancestral to a group or organisms with shared evolutionary history. Therefore, to be consistent with biological knowledge such phenotypes should be asserted only once for the ancestor, and then be inferred (down-propagated) for each descendant organism, and the taxa that these are members of. We can accomplish this by introducing the object properties *has_ancestor* (and its inverse *has_descendant*), which relates organisms to their ancestors, and *has_progenitor*, a subproperty of *has_member* that relates a taxon to the ancestral member of that taxon. For example, Siluriformes have lost scales, which we can express in the following way:

```
Individual: Siluriformes
Types: Clade, has_progenitor some
      (has_part only (not scale))
Facts: has_rank order
```

By defining the property chain ‘*subclade_of* o *has_progenitor* → *has_ancestor_to_members*’, we can straightforwardly express the set of all species with an ancestor that does not have scales as:

```
((has_rank value species) and
 (has_ancestor_to_members' some
 (has_part only (not scale))))’.
```

Traits are borne by, and observed for individual organisms. To facilitate free inter-

mingling of and reasoning across directly observed traits with those hypothesized as ancestral states, they should be expressed consistently as properties of organisms, rather than those of taxa. The individual-based approach we propose here accomplishes this in a straightforward manner.

Some of the presented issues can also be addressed within a class-based taxon model, for example by adding a level of indirection between the taxon class and the phenotype assertion. However, such approaches still make it difficult to appropriately propagate phenotypic data across levels of the taxonomy as shown above. Furthermore, assigning population-specific property values, such as geographic range or temporal extent, in a class-based approach suffers from similar problems, whereas it is straightforward in an individual-based modeling approach.

In conclusion, representing taxa as ontological classes presents an unrealistic model of organisms as instances of kinds, which creates a variety of problems for ontologically modeling observations of organismal properties. Embracing instead “population thinking” [2], and correspondingly modeling taxa, organisms, and their observed traits in a way that directly represents the historical individuality of evolutionary lineages could make integration and reasoning over comparative observational data in a fully OWL DL compatible framework not only much more straightforward, but also more consistent with our understanding of biological evolution.

References

1. Hull, D.L. A matter of individuality. *Philosophy of Science* 45(3), 335—360 (1978)
2. Mayr, E.. *Populations, species, and evolution*. Harvard University Press, Cambridge, MA (1970)

Phenoscape: Use Cases and Anatomy Ontology Requirements for Linking Evolutionary and Model Organism Phenotypes

Wasila Dahdul^{1,2}, James Balhoff^{2,3}, Hilmar Lapp², Peter Midford²,
Todd Vision^{2,3}, Monte Westerfield⁴, Paula Mabee¹

¹University of South Dakota, Vermillion, SD, USA; ²National Evolutionary Synthesis Center, Durham, USA;

³University of North Carolina, Chapel Hill, USA; ⁴University of Oregon, Eugene, USA

The naturally occurring phenotypes documented for groups of species in the systematics literature are recorded in free text which, although usefully precise and expressive, is not amenable to computational processing. This prevents its use for large-scale analysis or integration with the genetics knowledge that is available for single species in model organism databases [1,2]. Using ontologies, Phenoscape developed a knowledgebase (PhenoscapeKB; kb.phenoscape.org) that connects phenotypes for fish species with genetically characterized phenotypes for the zebrafish, *Danio rerio*, from the ZFIN database (zfin.org). Through ontology-based reasoning over expert knowledge in taxonomy, comparative anatomy and developmental genetics, the PhenoscapeKB is designed to enable the discovery of candidate genes for the natural diversity of phenotypes across taxa, and the aggregation of phenotypes across systematics studies to enable a global view of phenotype data available in large clades. These use cases necessitated the development of two multispecies anatomy ontologies: one for fishes (~30,000 species) and recently one for all vertebrates (~50,000 species), which will facilitate the expansion of Phenoscape to all vertebrates by connecting to other existing vertebrate ontologies and databases (amphibian, mouse, frog). These ontologies are required to fully represent the diversity of structures present in these extinct and extant species. It also required bridging across multiple scales of biological organization, from cells to anatomical systems. We discuss the implications that these requirements have for ontology design.

We developed the multispecies Teleost Anatomy Ontology (TAO) [3], with feedback from the ichthyological community, to represent the diversity of structures in teleost fishes and

for the phenotype curation of the fish systematics literature [4]. TAO was derived from, and is kept synchronized with, the species specific Zebrafish Anatomical Ontology (ZFA), which is used to describe phenotypes for the model organism *Danio rerio* in ZFIN. In expanding a single species ontology to one that is applicable to many species, existing term definitions were broadened to be universally applicable to all teleosts. New terms were added for structures present in teleosts but not in zebrafish, and this required the addition of grouping terms to TAO, which facilitate queries on similar structures for all fishes. Some of these grouping terms were unnecessary for ZFA (e.g., ‘tooth’ as a grouping term was not required in ZFA because zebrafish only have one type of tooth, represented by the term ‘ceratobranchial 5 tooth’). Some relationships between terms inherited from ZFA were removed from TAO because they were not applicable to all teleosts (e.g., the assertion ‘vertebra 1’ is_a ‘Weberian vertebra’ is valid for ZFA but not for TAO because not all teleost vertebra 1 are subtypes of Weberian vertebra). These taxonomically variable relationships have implications for deriving a single species ontology from a multispecies ontology, and vice versa, because relationships that might be required for a single species may be invalid in a multispecies context. We are exploring how to represent these relationships, possibly as annotations with *in_taxon* relationships to terms from a taxonomy ontology.

To initiate the expansion of Phenoscape to all vertebrates, we invited experts to a workshop to reevaluate and redefine existing skeletal terms (cells, tissues, development, anatomical structures) for their applicability across vertebrates, and to create terms for concepts not yet represented in existing

ontologies. This work resulted in the Vertebrate Anatomy Ontology (VAO), which will serve as a reference ontology for new and existing vertebrate subontologies. VAO accommodates the various ways that biologists classify bones and cartilages, e.g., as distinct elements and tissue types and based on developmental and locational criteria. Textual definitions are included for all terms, and an effort is currently underway to translate these text definitions into computable logical definitions. Logical definitions assist in ontology maintenance and error checking because a reasoner is used to automatically classify terms. Logically defined terms take the form of genus-differentia definitions, in which “X is a G that D” where X is the term being defined, G is the genus, and D is the differentia. In the example below, ‘replacement element’ (X) is a type of skeletal element (G) that participates in ‘replacement ossification’ (D; this term will be requested from the Gene Ontology and imported into VAO):

```
[Term]
id: VAO:0000135
name: replacement element
def: "Skeletal element that forms
    as a replacement or substitution
    of another element or tissue."
[VAO:curator]
intersection_of:
    VAO:0000128 ! skeletal element
intersection_of: participates_in
    VAO:0000140 ! replacement
    ossification
```

Synthesis and discovery in combined genetic, developmental and evolutionary data requires anatomy terms to be related across

scales of biological organization. The VAO enables this discovery by relating tissues to cells (e.g., ‘osteocyte’ part_of ‘cellular bone tissue’) and structures to processes (e.g., ‘replacement element’ participates_in ‘replacement ossification’). Putting in place these relationships to biological data at different scales will significantly increase the potential for discovery of candidate genes and taxa from queries on anatomical structure.

References

1. Mabee, P. M., G. Arratia, M. Coburn, M. Haendel, E. J. Hilton, J. G. Lundberg, R. L. Mayden, N. Rios, and M. Westerfield. 2007. Connecting evolutionary morphology to genomics using ontologies: A case study from Cypriniformes including zebrafish. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution* 308B:655-668.
2. Mabee, P., M. Ashburner, Q. Cronk, G. Gkoutos, M. Haendel, E. Segerdell, C. Mungall, and M. Westerfield. 2007. Phenotype ontologies: the bridge between genomics and evolution. *Trends in Ecology & Evolution* 22:345-350.
3. Dahdul, W. M., J. G. Lundberg, P. E. Midford, J. P. Balhoff, H. Lapp, T. J. Vision, M. A. Haendel, M. Westerfield, and P. M. Mabee. 2010. The Teleost Anatomy Ontology: anatomical representation for the genomics age. *Systematic Biology* 59:369-383.
4. Dahdul, W. M., J. P. Balhoff, J. Engeman, T. Grande, E. J. Hilton, C. Kothari, H. Lapp, J. G. Lundberg, P. E. Midford, T. J. Vision, M. Westerfield, and P. M. Mabee. 2010. Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. *PLoS ONE* 5:e10708.

Ontologies in the Fish Tank: Using the Zebrafish Anatomy Ontology with Other OBO Ontologies to Annotate Expression and Phenotype

Yvonne Bradford, Ceri Van Slyke

Zebrafish Model Organism Database (ZFIN), University of Oregon, Eugene, OR, USA

The Zebrafish Anatomy Ontology (ZFA) was created by the Zebrafish Model Organism Database (ZFIN) to facilitate curation of gene expression and phenotypes. The ZFA is an OBO Foundry ontology that contains 2648 terms and is 100% is_a parent complete. The ontology is updated with new terms, relationships and definitions on a regular basis, with pre-versions available prior to the update. (http://www.obofoundry.org/cgi-bin/detail.cgi?id=zebrafish_anatomy)

The ZFA is primarily used to curate expression and phenotype data and provides a canonical description of anatomical structures that facilitates cross-species comparisons. Curators use the ZFA to accurately capture gene expression data as they are described in the primary literature. ZFIN curators are able to compose gene expression statements in the curator interface using a ZFA term by itself, or using more specific post-composed terms created by coupling a ZFA term with a GO-cellular component (GO-CC) or Spatial Ontology (BSPO) term. Gene expression data comprised of annotations that utilize the ZFA, GO-CC, and BSPO, along with the Zebrafish Stage Ontology (ZFS), are provided on gene, genotype, and figure pages in ZFIN and are searchable and downloadable.

Additionally, ZFIN utilizes the ZFA to capture morphological phenotype statements that describe the effects of mutant or knocked-down gene products. ZFIN curators are able to represent entities (E) using the ZFA in conjunction with BSPO, GO-CC, GO-molecular function (GO-MF) and Cell type (CL) ontologies to produce post-composed statements. These entity statements are combined with Phenotypic quality (Q) terms (PATO) to create the E+Q[+E] phenotype description. The E+Q[+E] phenotype description, combined with the genotype, environment, and stages from the ZFS comprise a phenotype statement. Phenotype data can be accessed from the gene, morpholino, genomic feature, genotype, and figure pages and is available for download.

ZFIN is currently working toward adding the Mouse Pathology ontology (MPATH) and the Neuro Behavior Ontology (NBO) to be used in conjunction with ZFA and PATO for the curation of cancer and behavior phenotypes. ZFIN works to actively incorporate new ontologies to be used in the curatorial process that allow ZFIN curators to capture robust annotations that can be successfully used in cross-species comparisons by the wider community.

Doctoral & Post-Doctoral Consortium



ICBO

International Conference on Biomedical Ontology

July 27, 2011
Buffalo, New York, USA

One Empirical Study, Three Tests of Translation, Many Questions on Biomedical Ontologies: Limited Contribution of MeSH Terms to Effective Literature Searches on ‘Health-Related Values’

Mila Petrova

Egenis (ESRC Centre for Genomics in Society), University of Exeter, Devon, UK

Keywords: biomedical ontologies, ontologies of values, values in ontologies, search strategies, research-into-practice translation, evidence-based medicine, bio-psycho-social frameworks, boundary objects, promiscuous realism

This paper will raise three translation issues associated with biomedical ontologies: translation between biomedical ontologies and health-related ontologies informed by the social and behavioral sciences and the humanities; translation of ontologies into literature searching strategies in electronic bibliographic databases; and translation of intended applications of ontologies to the indexing of research publications into actual indexing practices.

The broad context in which these issues have been articulated is that of literature searches underpinning ‘research synthesis’ work (of which systematic reviews are the most influential type, but at least forty other methods have been identified in review publications). The narrow context is that of a study on developing search strategies for identifying publications on health-related values in electronic bibliographic databases. ‘Values’ was understood very broadly, to include issues such as patients’ and other stakeholders’ perceptions, preferences, beliefs, experiences, expectations, quality of life; issues of interdisciplinary communication; values underpinning diagnostic classifications, etc. The core element of the study was a word frequency analysis of datasets on Diabetes, Obesity, Dementia and Schizophrenia. These comprised 4,440 citations (2,449 “true positives” and 1,991 “false positives” for values contents; MEDLINE; Jan 2004 – Dec 2006) amounting to over a million-word textual corpus. Both text words and MeSH terms were analyzed. A 22-line values search filter was developed following principles of objective methods for search filter development. It had sensitivity and precision of 76.8% and 86.8% in the development dataset and between 47.1% and 70.1% (sensitivity) and 63.6% ÷ 82.6% (precision) in validation datasets. Text words came out as significantly more effective than MeSH terms relative to the desiderata for the filter (brief, of high precision and at least moderate sensitivity).

The work also showed unexpected patterns of assignment of terms to research papers for the purposes of MEDLINE indexing. For instance, it seems that “Attitude to Health” or “Health Knowledge, Attitudes, Practice” act as umbrella terms for almost anything “psychosocial”.

The first translation issue this study pointed to concerns the relationship between biomedical ontologies on the one hand and health-related ontologies coming from the social and behavioral sciences and the humanities on the other. Ontologies underpinning medical and health bibliographic databases were found to be either minimal in mapping the field of health-related values or aligned with concepts from the social and behavioral sciences and the humanities that do not translate easily into clinical practice and/or health policy concerns. Can the biomedical ontologies community justifiably treat this as an NMP (“not my problem”)? Could engagement with those “other” health-related ontologies enhance a biomedical ontology’s own capacity to support the translation of research into clinical practice?

The second translation issue identified concerns the capacity of biomedical and other health-related ontologies to contribute to effective literature searching strategies. In this study, controlled vocabulary terms had a limited role in an objectively developed search filter. Text words comprised by far the greater proportion of it (18 vs. 3 terms). Thus, current ontologies mapping the field of health-related values failed to translate into effective literature searching strategies. Is this a generic problem, of difficulties of translation between highly discriminative, fine-grained ontologies and coarser-grained user questions? Or is it a middle-ground one, for instance of translation of some ontologies into literature searches on broad, fuzzy topics? Or is it simply a very particular problem solvable through established means?

The third translation issue highlighted by this study concerns the actual, practically negotiated use of ontologies in the indexing of research publications. Findings suggested that certain patterns of assignment of controlled vocabulary terms to research papers may reflect historical, local and contingent practices rather than a reasonably direct matching of the explicit definition and scope of a term and the contents of a research paper. How can knowledge of such actual ‘workarounds’ help us improve the contents and structure of our ontologies? Are current feedback mechanisms between ontology developers and

different types of ontology users robust and varied enough?

To conclude, I will discuss whether the development and application of ontologies supporting research synthesis work or mapping fields of fuzzy concepts such as health-related values may be usefully informed by Susan Leigh Star’s concept of “boundary objects” and John Dupré’s ideas on pluralism and “promiscuous realism”. I will also suggest some priorities for empirical research on the actual use of biomedical ontologies in terms of translation issues.

The Age of Data-Driven Medicine: Mining the Electronic Health Record

Paea LePendu, Mark A. Musen, Nigam H. Shah

Stanford Center for Biomedical Informatics Research, Stanford, CA, USA

With the availability of tools for automated coding of unstructured text using natural language processing, the existence of over 250 biomedical ontologies, and the increasing access to large volumes of electronic medical data, it is possible to apply data-mining techniques to the large amounts of unstructured data available in medicine and health care. For example, by computationally encoding the free-text narrative—comprising roughly 80% of the clinical electronic medical data—it may be possible to test drug safety signals in an active manner. We describe the application of NCBO Annotation tools to process clinical text and the mining of the resulting annotations to compute the odds ratio of having a myocardial infarction on taking Vioxx for Rheumatoid arthritis. Our preliminary results demonstrate that it is possible to apply annotation analysis methods for testing hypotheses about drug safety using electronic medical records.

Given recent advances in detecting drug

safety signals from spontaneous reporting systems, it becomes crucial to develop methods of searching for, testing, and applying these signals throughout the electronic health record so as to realize their benefits on new patients before an adverse event occurs. We hypothesize that using ontology based approaches, analogous to enrichment analysis, can fill this gap.

One of our recent successes has been the considerable improvement of the NCBO Annotation tools. We have optimized the system for both speed and space so that it can extract clinical concepts from the textual reports of the entire Stanford Clinical Data Warehouse—nearly 10 million surgical pathology, radiology, and general transcription reports—overnight, on a single machine. Furthermore, we have added negation detection as well as extended concept recognition capabilities, such as morpheme-based matching. We are highly encouraged by our preliminary results in detecting the Vioxx risk signal from that data.

Quality of Care Domain Modeling in Cancer: A Semantic Approach

Sina Madani^{1,2}, Dean F. Sittig¹, Parsa Mirhaji¹, Kim Dunn¹

¹University of Texas, Health Science Center, School of Biomedical Informatics, Houston, TX, USA

²University of Texas, MD Anderson Cancer Center, Houston, TX, USA

Abstract. There is an increasing demand from heterogeneous organizations to collect and report healthcare quality metrics. To create an overarching model of quality that captures all stakeholders' perspectives of care and compare quality of care metrics consistently, clinical data elements should be modeled and represented unambiguously. We propose to use semantic web technologies in the domain of cancer care to build such a harmonized model and generate explicit, consistent, and comparable reports.

1 Background and Significance

The Institute of Medicine reports a growing demand in recent years for quality improvement within the healthcare industry [1-4]. In response, numerous organizations have been involved in the development of *quality measurement metrics*. However, the quality metrics development is subjective in nature [5] and competing interests exist among developer organizations. As a result, conflicting data definitions from such organizations shift the burden of accurate and reliable metrics extraction and reporting upon healthcare providers [6-8]. Furthermore, manual abstraction of quality metrics [8,9], diverse implementation of Electronic Health Record (EHR) Systems [8,10], and the lack of standards for integration across disparate clinical and research data sources [11] deepens the complexity of *consistent*, *valid*, *explicit*, and *comparable* quality measurement reporting task within healthcare provider organizations.

A successful model for healthcare quality measurement must represent an integrated and comprehensible view for all stakeholders from payers to providers to patients [12]. In order to construct such an overarching model, concepts should be defined explicitly, such that heterogeneous information from different sources can be reliably mapped and compared based on those concepts. While a reference information model, like the proposed National Quality Forum (NQF) Data Model [13] or eMeasures [8], can be used for deriving a

syntactic data model [14], it does not represent such a shared and comparable data semantics [15] for *harmonized representation* of heterogeneous schemas [14, 16]. In addition, neither is there a well defined interface between such information models and the EHR systems [17] nor can cancer quality concepts be represented solely by such a complex syntactical standard [18, 19]. Hence, quality metrics developed by diverse organizations, as well as provider's own internal metrics, cannot be modeled exclusively, compared explicitly, and *represented unambiguously* by standards such as the proposed reference information models [19].

2 Research Method and Design

In the first phase of this study, we will explore all existing quality measurement metrics and their definitions in the domain of breast cancer (eight metrics). We intend to use a formal method to construct an unambiguous semantic nomenclature from explored definitions. The method will explicitly define all concepts, show mappings among concepts, validate the model, normalize attributes, bind concepts into standard terminologies [16], provide a holistic view, and facilitate federated query functionalities [20]. *In the second phase* we will create a comprehensive conceptualized model using a standard semantic specification [21] for harmonized representation of concepts. The build process will include formal definitions of concepts and their relationships from the explored components in the previous phase. *In the third*

phase we will perform a series of semantic queries on a target group of quality metrics instances from federated data and compare the results against current query techniques for *functional* validation of the model. Similar comparison will be made between components of the model and the existing manual collection of the metrics by domain experts for validating completeness of the *domain coverage*. Finally, available semantic rule engines will be used for *structural* validation of the model. The host institution for this proposal, MD Anderson Cancer Center, is the largest freestanding cancer center in the world. There were 105,000 patients who visited MD Anderson in 2010 [22], thus providing the primary investigator with a large amount of data for validation of the proposed model.

The specific aims of this proposal are the following:

1. Explore existing quality of care measurement models in breast cancer. Identify components, relationships, and stakeholders within and across models.

2. Create a conceptual model of the quality metrics in the breast cancer care domain. Formally define model components and relationships. Propose a comparative mapping across various metrics and their components.

3. Evaluate the model. Validate structure, function, and domain coverage of the model.

References

- Kohn, L.T., Corrigan, J.M., Donaldson, M.S.: To err is human: building a safer health system. A report of the Committee on Quality of Health Care in America, Institute of Medicine. (2000)
- Institute of Medicine. Committee on Identifying and Preventing Medication Errors. (2006)
- Institute of Medicine. Committee on Redesigning Health Insurance Performance Measures Payment and Performance Improvement Programs: Rewarding provider performance: Aligning incentives in Medicare. (2007)
- Institute of Medicine. Committee on Quality of Health Care in America: Crossing the quality chasm. (2001)
- Miller, R.D.: Miller's anesthesia. Churchill Livingstone/Elsevier, Philadelphia, PA (2010)
- Institute of Medicine. Committee on Redesigning Health Insurance Performance Measures Payment and Performance Improvement Programs: Performance measurement: accelerating improvement. National Academies Press (2006)
- Tang, P.C., Ralston, M., Arrigotti, M.F., *et al.* Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an EHR system. J Am Med Inform Assoc 14 (2007)
- Velamuri, S.: QRDA – Technology Overview and Lessons Learned. J Health Inf Manag 24, 2010
- US Department of Health and Human Services, <http://www.ncvhs.hhs.gov/080128lt.pdf>
- McDonald, C.J.: The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 4, 213-221 (1997)
- Chong, Q., Marwadi, A., Supekar, K., Lee, Y.: Ontology based metadata management in medical domains. J Research and Practice in Information Technology 35, 139-154 (2003)
- Spinks, T.E., Walters, R., Feeley, T.W., *et al.*: Improving cancer care through public reporting of meaningful quality measures. Health Aff (Millwood) 30, 664-672 (2011)
- National Quality Forum, http://www.qualityforum.org/Projects/h/QDS_Model/Quality_Data_Model.aspx
- Carlson, D., Farkash, A., Timm, J.T.: A model-driven approach for biomedical data integration. Stud Health Technol Inform 160, 1164-68 (2010)
- Smith, B., Ceusters, W.: HL7 RIM: an incoherent standard. Stud Health Technol Inform 124, 133-138 (2006)
- Bianchi, S., Burla, A., Conti, C., Farkash, A., Kent, C., Maman, Y., Shabo, A.: Biomedical data integration. Conf Proc IEEE Eng Med Biol Soc 2009, 4654-4657 (2009)
- Ferranti, J.M., Musser, R.C., Kawamoto, K., *et al.* The clinical document architecture and the continuity of care record: a critical analysis. J Am Med Inform Assoc 13, 245-252 (2006)
- Pisanelli, D.M., Gangemi, A.: If ontology is the solution, what is the problem? Stud Health Technol Inform 102, 1-19 (2004)
- Muir, E., <https://interfaceware.fogbugz.com/default.asp?W252>
- Kamal, J., Borlawsky, T., Payne, P.R.: Development of an ontology-anchored data warehouse meta-model. AMIA Annu Symp Proc (2007)
- McGuinness, D.L., Van Harmelen, F.: OWL web ontology language overview. (2004)
- MD Anderson Cancer Center, <http://www.mdanderson.org/about-us/facts-and-history/fact-sheet/index.html>

Philosophy, Ontology, and Scientific Explanation

James A. Overton

The University of Western Ontario, Canada

Explanation is one of the goals of science. One of the goals of the philosophy of science is to account for the logic of scientific explanation. But advances in informatics technologies are changing science in important ways. In my doctoral research I am developing a philosophical account of scientific explanation with a focus on the role of informatics in science. I consider how scientist are using, and could best be using, informatics technologies to express, test, and communicate their scientific explanations.

Biomedical ontologies are an excellent example of the productive intersection of science, informatics, and philosophy. Using ontologies, scientists are able to coordinate on systems of terminology, encode their data for better reuse and sharing, and search and analyze that data more effectively [2,16]. My goal is an account of scientific explanation that builds on these successes.

Philosophers of science have developed a number of accounts of explanation, from deductive-nomological [8,7] to unification [6,9] and pragmatic explanation [17], and from causal-mechanical [15,5] to mechanistic [11,3], interventionist [18], and model-based explanation [4]. These diverse accounts have many insights, but none of them is a good fit for all of the many sorts of scientific explanation. None of these accounts were designed with informatics systems in mind, although deductive-nomological and unification accounts have been adopted in artificial intelligence research on explanation [10, p.68].

I begin by distinguishing the sorts of things that are explained and do the explaining in scientific articles: entities, data, kinds, models, and theoretical considerations. By “entities” I mean concrete particular things; data are records of measurements and classifications of entities; kinds are similar to the universals that ontologies describe; and by “models” I mean (roughly) abstract systems for drawing inferences, such as equations, algorithms, and perhaps some narratives. I support these distinctions with textual evidence from

scientific journals.

One important type of explanation is model-kind explanation, where aspects of a model explain aspects of a kind. The kind could be a biochemical pathway, such as glycolysis, while the model is a network of molecule kinds and reaction kinds, including glucose, ATP, and phosphorylation [12]. Ontologist use decidable fragments of first order logic to reason over kinds and their relations. Scientific modelling usually requires more powerful logics and mathematics. I am exploring the use of typed programming languages for expressing such models.

This theoretical work on scientific explanation is supplemented by my applied research into the use of ontologies for reporting in radiology [1,14,13]. My hope is that a better understanding of scientific explanation and the role of informatics in science will lead to better explanations, and to better informatics systems for building, testing, and sharing explanations.

References

1. Arp, R., Romagnoli, C., Chhem, R.K., Overton, J.A.: Radiological and Biomedical Knowledge Integration: The Ontological Way. In: Chhem, R.K., Hibbert, K.M., Van Deven, T. (eds.) *Radiology Education: The Scholarship of Teaching and Learning*, pp. 87–104. Springer, Heidelberg (2008)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: *Gene Ontology: Tool for the Unification of Biology*. *Nature Genetics* 25(1), 25–29 (2000)
3. Bechtel, W., Abrahamsen, A.: *Explanation: A Mechanist Alternative*. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36, 421–441 (2005)
4. Bokulich, A.: *How Scientific Models Can Explain*. *Synthese* 180(1), 33–45 (2009)
5. Dowe, P.: *Physical Causation*. Cambridge

- University Press, Cambridge (2000)
6. Friedman, M.: Explanation and Scientific Understanding. *The Journal of Philosophy* 71(1), 5–19 (1974)
 7. Hempel, C.G.: Aspects of Scientific Explanation. In: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, pp. 331–496. Free Press, New York (1965)
 8. Hempel, C.G., Oppenheim, P.: Studies in the Logic of Explanation. *Philosophy of Science* 15(2), 135–175 (1948)
 9. Kitcher, P.: Explanatory Unification and the Causal Structure of the World. In: Kitcher, P., Salmon, W.C. (eds.) *Scientific Explanation*, pp. 410–505. Minnesota Studies in the Philosophy of Science, University of Minnesota Press, Minneapolis (1989)
 10. Leake, D.B.: *Evaluating Explanations: A Content Theory*. Lawrence Erlbaum Associates, Hillsdale, NJ (1992)
 11. Machamer, P.K., Darden, L., Craver, C.F.: Thinking About Mechanisms. *Philosophy of Science* 67(1), 1–25 (2000)
 12. Overton, J.A.: Mechanisms, Types, and Abstractions. *Proceedings of the Philosophy of Science Association* 2010 (Forthcoming)
 13. Overton, J.A., Romagnoli, C., Chhem, R.K.: Open Biomedical Ontologies Applied to Prostate Cancer. *Applied Ontology* 6(1), 35–51 (2011)
 14. Romagnoli, C., Overton, J.A., Chhem, R.K.: Philosophy in Radiology: The Ontological Challenge. In: Van Deven, T., Hibbert, K.M., Chhem, R.K. (eds.) *The Practice of Radiology Education*, pp. 239–248. Springer, New York (2010)
 15. Salmon, W.C.: *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton (1984)
 16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., the OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nature Biotechnology* 25(11), 1251–1255 (2007)
 17. van Fraassen, B.C.: *The Scientific Image*. Clarendon Press, Oxford (1980)
 18. Woodward, J.: *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford (2003)

Refining the Ontology for Glucose Metabolism Disorders

Yu Lin

Center for Computational Medicine and Bioinformatics, Unit of Laboratory Animal Medicine, and
Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI, USA

Abstract. The Ontology for Glucose Metabolism Disorders (OGMD) was recently refined to meet two purposes: 1) contribution to Disease Ontology (DO); 2) to allow application to a social network study for Diabetes Mellitus researchers. It has been refined by applying the OBO Foundry principles.

The Ontology for Glucose Metabolism Disorders (OGMD) was developed four years ago mainly for supporting the translational study on Diabetes Mellitus. It is applied with the ontology of Geographical regions (OGR) and the Ontology of Genetic Susceptibility Factor (OGSF) for describing genetic susceptibility factors to Diabetes Mellitus [1]. Not only limited to the usage of its original purpose, OGMD as a terminology of disease names has gained attentions to informatics researchers in the field of Diabetes Mellitus (DM). For example, a group in Thailand who developed a Type II Diabetes Mellitus Clinical Support System has been reviewing OGMD in their study [2]. Another group in University of Michigan will use OGMD to analyze the social network of researchers working on Glucose Metabolism Disorders. Regarding users' requirement and the requirement of ontology's evolution by itself, OGMD has been refined recently and gained insights on the issue of ontology sharing.

There are two purposes to refine OGMD. One purpose is that OGMD contributes to Disease Ontology by providing it with disease terms. The other purpose is to serve the project of analyzing the social network of Diabetes related researchers located inside University of Michigan.

Since the initiation of OBO Foundry, the community has built up solid rules for developing OBO Foundry ontologies. It is necessary to follow the rules of OBO Foundry since OGMD has been built before those rules were released. An important change made in OGMD is to follow the OBO ID policy. Another change is that Relation Ontology (RO) has

been imported into OGMD. Since DO is in obo format, OGMD has been developed as an OBO file as well.

There are 131 terms in OGMD. OGMD is available both in OWL and OBO formats. The classification of OGMD has been aligned with the related terms in DO. Among 83 OGMD terms that are not included in DO, 25 OGMD terms have been submitted to DO. OWL file is ready for the social network project application.

New classes have been developed, such as Diabetes complications, Diabetes as a complication of another disease. Using the objective relation *has_disposition* and *disposition_of* in RO (proposed version), the statements in OGMD: "Diabetes *has_disposition* Diabetes complications"; "Diabetes is a complication of other disease *disposition_of* Other disease" were established. Applying OGMD to the social networking project is on-going, and this application will give evaluation and validation to OGMD in the near future. It may enlarge the potential of sharing ontology by using only the minimum core relations and core terms within a domain.

References

1. Lin Y, Sakamoto N: Ontology driven modeling for the knowledge of genetic susceptibility to disease. *Kobe J Med Sci* **55**(3):E53-66.(2009).
2. Chalortham N, Buranarach M, Supnithi T: Ontology Development for Type II Diabetes Mellitus Clinical Support System. *Proc. of the 4th International Conference on Knowledge, Information and Creativity Support Systems (KICSS2009)*. (2009).

A Case Study in Using ZFA and PATO for Describing Histological Phenotypes in the Larval Zebrafish

Brian Canada¹, Georgia Thomas², Timothy Cooper², Keith Cheng²

¹University of South Carolina Beaufort, Bluffton, SC, USA

²Penn State College of Medicine, Hershey, PA, USA

When used in conjunction with an appropriate ontology of quality-bearing entity terms, such as the Zebrafish Anatomy and Development (ZFA) Ontology [1], the Phenotype and Trait Ontology (PATO) [2] permits the description of phenotypic qualities by means of a bipartite “entity-quality” (EQ) data structure. Potentially, the use of these ontologies can mitigate ambiguity in phenotypic interpretation that might otherwise occur using free-text, plain language descriptions. For example, the phrase “fin edges look ratty” ascribed to the stereomicroscopic characterization of the zebrafish insertional mutant *parvaa* is intended to describe defects in the epithelial tissue of the fin, but this might not be apparent to someone unfamiliar with such colloquialisms. A possible coercion of this phrase to the ZFA-PATO format could be expressed as “surface structure: abnormal,” as it is currently described in the Zebrafish Information Network (ZFIN) database [3]. However, some precision is lost in the conversion – the researcher does not immediately know *which* surface structure is abnormal, nor does he or she know *how* abnormal it is.

Here, we will present our findings from a case study in which we have explored the use of ZFA and PATO for the description of histological phenotypes in larval zebrafish. Like the *parvaa* example, we have found that ZFA-PATO is imprecise with respect to tissue and cell type affected. Many zebrafish mutants can only be distinguished at microscopic levels of histological detail that current ZFA-PATO term combinations cannot fully describe. Based on our experience, we advocate the adoption of more highly specific terms into ZFA and PATO,

perhaps based on terms “borrowed” from other model organism ontologies that currently accommodate the desired level of precision. We also propose that ontology terms (or at least their definitions) reflect the assay and resolution used (e.g., “dissecting microscope”) so that researchers can easily recognize the limits of current knowledge in phenotype databases such as ZFIN. In addition, we demonstrate that even further precision (such as for describing the *extent* of abnormality) can be achieved by employing the under-utilized *modifier* term in the PATO data structure, which can allow for quantitative measurements as well as semi-quantitative “grades” not unlike those used by pathologists in characterizing disease progression. Enabling such precision gains may be the only way to describe phenotypes at the level of “granularity” needed for the forthcoming Zebrafish Phenome Project [4].

References

1. Ontology Detail: Zebrafish anatomy and development,
[http://www.obofoundry.org/cgi-bin/detail.cgi?id=zebrafish anatomy](http://www.obofoundry.org/cgi-bin/detail.cgi?id=zebrafish%20anatomy)
2. Ontology Detail: Phenotypic quality,
<http://www.obofoundry.org/cgi-bin/detail.cgi?id=quality>
3. ZFIN: Figure: Phenotype Annotation (1994-2006) for (hi1320),
<http://zfin.org/cgi-bin/webdriver?MIval=aa-fxfigureview.apg&OID=ZDB-FIG-070117-324>
4. Zebrafish Phenome Project 2010 Meeting,
<http://www.blsmeetings.net/zebrafish/>

Epistemological Issues in Information Organization Instruments: Ontologies and Health Information Models

André Queiroz Andrade, Maurício Barcellos Almeida

Escola de Ciência da Informação - Universidade Federal de Minas Gerais, Pampulha - Belo Horizonte – MG – Brazil

Abstract. This paper describes the research for a methodology to represent health information in medical records, using realist ontologies and information models.

Keywords: Realist ontologies, information model, electronic health record

1 Introduction

Full use and sharing of medical data contained in health records depends on the capacity to semantically represent messages, store messages, receive queries and answer them. Ontologies are an alternative, since they rigorously define basic properties of the entities and necessary criteria required to instantiate a type. Particularly, the Ontological Realism seems capable of promoting consensus and internal coherence by representing reality according to a philosophical realist perspective, using science to get closer to the truth [1, 2]. However, medical records contain many terms and expressions that lack a referent in reality, but are nevertheless important for medical communication and information use [3, 4]. On other hand, information models create a language based information structure that allows complete information representation, though lacking consistency, capability to make inferences and internal coherence.

2 Objectives

We aim to propose and practically validate a methodology for information representation of health information present in real medical records, through the complementary use of ontologies and information models.

3 Methodology

The methodology is divided in two parts: methodology proposition; methodology validation. The proposition will be made as follows:

1. Requirement analysis: We will evaluate technical and logical requirements for medical information representation;
2. Selection of real medical records: We will select persistent documents contained in medical records (physician notes, discharge and admission reports, lab test results) written in natural language and de-identify the records;
3. Create OpenEHR records: We will create new records according to the OpenEHR model, transforming the natural language text into structured information. E.g. “Patient complains of crushing left chest pain” → Cluster – Pain Symptom; Name of location: “Precordial area”;
4. Analysis and classification of information: We will try to map the OpenEHR information items to Realist Ontologies, such as those contained or candidate to OBO Foundry inclusion. E.g. Name of location: “Precordial area” → FMA.Precordium;
5. Those terms that can’t be mapped will be used as parameter to identify boundaries between ontologies and information models, and to propose a to represent information which is not suitable to the realist approach while retaining the capacity to classify and manipulate information. Particularly, we will explore the use of information ontologies to represent the information models [5].
6. Formulate competency questions: We will create queries that must be answered by

information created using the methodology, in order to validate it.

The validation phase will be made by:

1. Random real records selection: Some records will be randomly selected – the exact number is yet to be determined.
2. Record representation using the proposed methodology coupled with ontological realist representation;
3. Critical evaluation using competency questions: We will evaluate the methodology according to some objective criteria, to demonstrate or suggest its efficacy.

References

1. Smith, B.: From concepts to clinical reality: An essay on the benchmarking of biomedical terminologies. *Journal of Biomedical Informatics* 39, 288-298 (2006)
2. Smith, B., Ceusters, W.: Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied Ontology* 5, 139–188 (2010)
3. Bodenreider, O., Smith, B., Burgun, A.: The Ontology-Epistemology Divide: A Case Study in Medical Terminology. In: 3rd Conference on Formal Ontology in Information Systems. (2004)
4. Schulz, S., Stenzhorn, H., Boeker, M., Smith, B.: Strengths and limitations of formal ontologies in the biomedical domain. *RECIIS Rev Electron Comun Inf Inov Saude* 3, 31-45 (2009)
5. Schulz, S., Schober, D., Daniel, C., Jaulent, M.C.: Bridging the semantics gap between terminologies, ontologies, and information models. *Stud Health Technol Inform* 160, 1000-1004 (2010)

The Foundational Model of Neuroanatomy Ontology: An Ontology Framework to Support Neuroanatomical Data Integration

B. Nolan Nichols¹, Jose L.V. Mejino Jr.², James F. Brinkley^{1,2}

¹University of Washington, Biomedical and Health Informatics, Seattle, WA, USA

²University of Washington, Biological Structure, Seattle, WA, USA

Keywords: Biomedical Informatics, Ontology, Semantic Web, Neuroinformatics

A basic requirement for facilitating the integration and analysis of neuroscience data from diverse sources is a well-structured ontology that can support multi-modal and multi-scale neuroscience applications. Here we show how a reference ontology, the Foundational Model of Anatomy (FMA) Ontology [1], meets data integration needs by applying a disciplined modeling approach to create an application ontology for neuroscience called the Foundational Model of Neuroanatomy (FMN) Ontology. The FMN provides a structural semantic framework useful for correlating and aligning disparate views of neuroanatomy, such as those present in the different ontologies and controlled terminologies used in neuroinformatics. The labels used in brain atlases are a form of controlled terminology that are frequently used to label neuroimaging result data sets, which creates an opportunity to integrate annotated data based on anatomical and structural relationships [2]. However, different brain atlases use distinct parcellation schemes that represent neuroanatomical structures and regions at different levels of granularity and also model different properties of neuroanatomical entities (e.g., cortical gray matter vs. white matter connectivity). In order to reconcile the labels used in several brain atlases, we enriched the FMA with additional neuroanatomical classes and properties to capture the semantics implicitly expressed in each brain atlas terminology.

By capturing the semantics implied in brain atlases, the FMN enables the correlation of annotated neuroimaging data sets that measure distinct or overlapping aspects of brain structure and function. After mapping

the anatomical labels used by common brain atlases (e.g., AAL, FreeSurfer, Talairach, and JHU-DTI-81) to FMA classes, we can now use the rich spatio-structural relation network (e.g. parthood, connectivity) of the FMA to determine the precise relationships between the structures represented both within and across different brain atlases. We specifically elaborated on parthood relationships to establish a granularity association between gray matter regions (e.g. gray matter of precentral gyrus *part_of* precentral gyrus) and within white matter tracts (arcuate fasciculus *part_of* superior longitudinal fasciculus). In addition we represented the neural connectivity relationships between gray and white matter structures.

To formalize structural connectivity in the FMA, we developed a set of definitions to disambiguate and clarify the terminologies used for describing the types of connectivity relationships that exist between gray and white matter structures (e.g. *projects_to*, *projects_from*). The *projects_to* property is a connectivity relation where individual axons comprising a fiber tract originating from one or more brain regions synapse with neurites or somas of a collection of neurons located in one or more brain regions. The *projects_from* property is a connectivity relation where individual axons comprising a fiber tract are parts of a collection of neurons located in one or more brain regions. The extended FMA can cross-correlate atlases using both part and connectivity relations, allowing measurements of white matter structures annotated with one atlas to be semantically integrated with measurements of gray matter structures annotated with a different atlas (Figure 1).

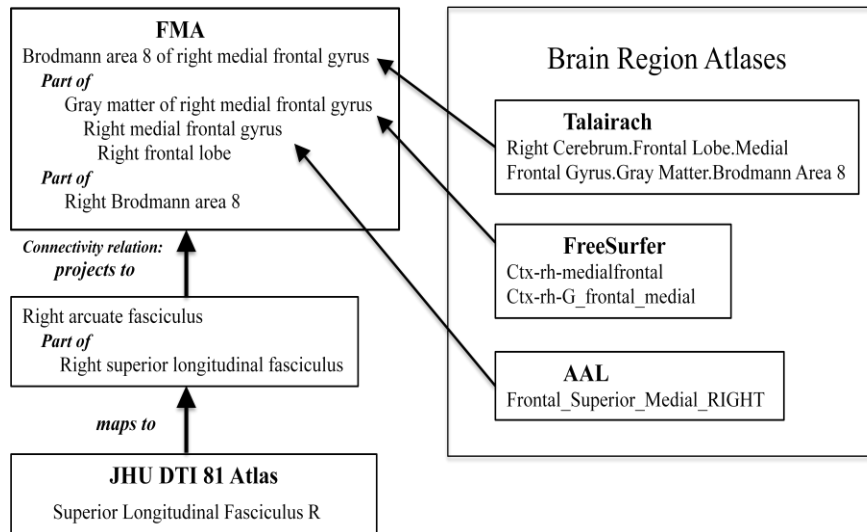


Figure 1. Cross-correlation of white matter tract with the different gray matter regions.

References

1. Rosse, C., Mejino, J.L.V.: A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy. *J. Bio. Info.* 36, 478–500 (2003)
2. Turner, J.A., Mejino, J.L.V., Brinkley, J.F., Detwiler, L.T., Lee, H.J., Martone, M.E., Rubin, D.L.: Application of neuroanatomical ontologies for neuroimaging data annotation. *Front. Neuroinform.* 4(10), 1–12 (2010).

Workflows and Framework for Nutritional/Metabolic Phenotype/Genotype and Foods-for-Health Knowledge Integration

Matthew Lange^{1,2}, J. Bruce German¹, Jim Kaput²

¹Department of Food Science, University of California at Davis, CA, USA

²Division of Personalized Nutrition and Medicine, US Food and Drug Administration, Jefferson, AR, USA

Keywords: food, diet, health, gene, genetic, genome, genomics, metabolic, metabolism, metabolome, metabolomic, nutrition, nutritional, phenotype, ontology, public health, vocabulary

Metabolic syndrome affects more than one-third of the adults in the United States, and is known to be influenced by diet. When this basic statistic, and its public health implications, is taken into account with other diet-related illnesses, the importance of considering how different foods can improve or detract from human health could not be more clear. Yet, as food and nutrition research attempt to address their next set of challenges (post essential nutrient discovery) i.e. health improvement, performance enhancement, quality of life enrichment and prevention of diseases related to diet and food choices – it finds itself radically under-served from informatics and knowledge management perspectives.

Many life-science disciplines relating to health, nutrition, food, and agriculture have developed databases, thesauri, and/or ontologies to capture their domain knowledge. Unfortunately, the language used to describe substantially similar (even logically equivalent) concepts is often different between information systems. High throughput and omics technologies that are expanding both the amount and heterogeneity of available information – will only muddy the water unless a common unifying infrastructure is developed.

Increasing the future value of agriculture while decreasing the future cost of healthcare therefore, will depend on creating a process for generating ontological commonalities that stretch from the characterization of food production and processing; their impacts on organoleptic and nutrient characteristics of the food; food preferences and eating habits; health

characteristics driven by food choices and reflected by metabolic phenotype; the up/down-regulated pathways and single nucleotide polymorphisms that give rise to these phenotypes, and the frequencies and degree to which these nutritional phenotypes and genotypes are correlated with each other, as well as the extent to which they exist in populations.

Presented here is a suite of workflows for integrating nutritional phenotype/genotype experiment information, look-up libraries for inter-knowledge-domain queries, and an illustration of a unified framework for identifying and integrating the most important foods-for-health informata. A common system of language that describes the food – phenotype – health/performance continuum, and is shared by all relevant life-science disciplines, will provide immediate benefits in terms of:

- better understanding of the complex causes of diseases in humans,
- increased health-claims transparency and improved regulatory efficiency,
- predictive functions for individualized foods and diets,
- promoting health and well-being including people who are already healthy.

Next steps include the building of ontological maps across the identified knowledge domains that will enable cohesive semantic markup and robust food-for-health querying.

Applications for a Translational Biomedical Ontology Model

Robert Yao, Graciela Gonzalez

Arizona State University

1 Introduction

In the past two centuries, the amount of scientific (especially medical) information has exponentially increased. As more became discovered, new paradigms required novel vocabularies and terms. This created new specialties and subspecialties of scientists and health professionals. They became subcultures with different languages. For instance, in general medicine, 'idiopathic' means: unknown cause; while in neurology, it means: presumed genetic.

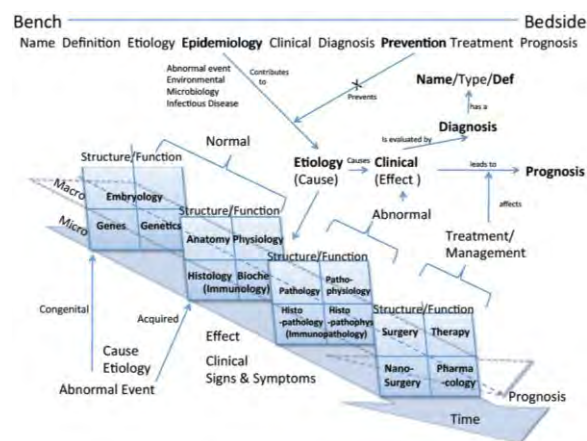
2 Problem

Sub-specialization coupled with the inundation of data has been problematic. First the current infrastructure does not take into account semantic similarities between these specialties. Second, because a coherent semantic infrastructure is missing, experts in those fields struggle to adequately understand disease and define syndromes. This is especially the case in neurology. For instance, inferences made from new findings could change the fundamental understanding of Alzheimer's disease. This in turn affects the kinds of questions researchers need to ask. Additionally, the criteria with which to define syndromes in Epilepsy are subject to change. This affects the ability for clinical researchers to locate patients that fit their criteria. In other words, this all creates an inefficient scenario where too much information builds up and keeping up with the literature proves difficult. Thus, understanding of the disease is incomplete, which then adversely affects a researcher's ability to uniquely research an unanswered question and/or to locate the right subjects for clinical research.

3 Background

Techniques in information retrieval like machine learning have been proven to be

effective at clustering and classifying data. However, without specific objectives or rules pertinent to the domain, the use of machine learning alone has run into various limitations in medicine. Ontologies which specify how concepts are represented are attempting to semantically standardize how information is handled and provide rules for the categorization of terms and defining their relationships. By connecting concepts and relationships to specific instances and information, a knowledgebase is instantiated using the ontology as its schema. While various biological and medical ontologies exist, finding the correct mapping combinations to answer clinical questions is still a major challenge.



4 Proposal

In an effort to solve this dilemma on several fronts, this work proposes a translational biomedical ontology for human disease. Rather than classifying where a disease name fits in a hierarchy (like most disease ontologies), it provides a robust ontologic infrastructure with which to help define any specific disease and its progression through various states that develop through time. Additionally, it does this from the perspective of a basic scientist to medical doctor to a clinical researcher. Thus it potentially aids

the organization, understanding, and transfer of information from bench to bedside and back.

5 Current Projects

In one use case, part of the ontology is being used on a corpus to train and test a machine learning algorithm that recognizes and semantically tags sentences in scientific and medical text about Alzheimer's Disease. The purpose of this is to build an algorithm for a system that automatically updates its own knowledgebase. It can scan sentences and automatically classify information about a disease and thus update that information in a knowledgebase. The growing knowledgebase is then used as the training and testing corpus and helps to further refine the algorithm for future scans. The knowledgebase also serves as a semantically searchable database that can be filtered for information that answers a specific question or helps someone to better understand the disease or the direction that future research must take.

In another use case, the terms and relationships present in the ontology itself are currently being explored in the definition of Epileptic seizures and syndromes. This is to be used with clinical data in i2b2 for the purpose of identifying subjects for clinical studies and trials. Information gathered at this stage will be mined, semantically tagged, and added to the knowledgebase. This will add to the corpus used to the machine learning model and hopefully result in more accurate tagging which will lead to better understanding.

6 Methods

As a first step towards proving this, efforts to automatically populate instances were pursued. The ontology was used in a machine learning classifier to semantically tag sentences automatically. In this project, over 1000 diseases were manually reviewed and semantically tagged. Categories that define disease in the ontology

were chosen to be the semantic tags: Etiology, Epidemiology, Clinical, Diagnosis (Tests), Prevention, Treatment, and Prognosis. For the purpose of this experiment, sentences pertaining to Alzheimer's disease in Harrison's and Cecil's Textbooks of Internal Medicine and Robbin's Pathology were manually curated and tagged to produce a corpus. A program was created to do a 10x10 cross validation of the corpus using Support Vector Machine and Logistic Regression classifiers in WEKA to test and train a model for automatic classification.

7 Results

The experiment is ongoing and results preliminary. Precision, recall, and f-measure of the model indicate high precision (85-100%) and low recall (7-20%) which adversely affects the f-measure. The low recall can be improved by the removal of stop words, stemming, adding more tagged sentences to the corpus. Perhaps active learning, or including human feedback in the process, may improve results. All of these can be attempted to maximize f-measure. Given that clinical text tends to focus on discussing epidemiology, clinical signs and symptoms, diagnostic tests and procedures, and prognosis of disease more than etiology, it is not surprising that etiology had the lowest precision. This may be improved further by adding more basic science text into the corpus.

8 Conclusion

The high precision indicates that while much work still needs to be done to improve recall, the use of learning algorithms to automatically tag ontology terms is a viable process. This provides the means to formalize a difficult problem that needs further testing. The next step is to expand the use of the other semantic terms and to determine if the tagged sentences from these are able to be searched for answers to biomedical questions.

	Etiology		Epidemiology		Clinical		Diagnosis		Prevention		Treatment		Prognosis	
	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR	SVM	LR
Precision	86%	87%	95%	98%	98%	98%	92%	91%	100%	99%	99%	91%	92%	99%
Recall	20%	20%	11%	11%	13%	11%	21%	21%	4%	4%	6%	8%	12%	12%
F-measure	32%	32%	19%	20%	22%	20%	34%	35%	8%	9%	12%	14%	22%	22%

AUTHOR INDEX

- Abdulla, Tariq 47*
 Abeyruwan, Saminda 209
 Adamusiak, Tomasz 19
 Ahmed, Mansoor 260
 Ai, Jiye 227, 381*
 Alexander, Paul 292
 Almeida, Mauricio Barcellos 227*, 381, 442
 Alpkocak, Adil 159
 Anderson, David 263
 Andrade, André Queiroz de 381, 442*
 Arighi, Cecilia N. 285*
 Arney, David 335*
 Athey, Brian D. 25

 Balci, Pinar 159
 Balhoff, James P. 230*, 426*, 428
 Bandrowski, Anita 263, 349
 Bastos, Hugo 300
 Batchelor, Colin 201
 Bello, Susan M. 231*
 Bertone, Matthew A. 422
 Berzell, Martin 143
 Blake, Judith A. 231, 259, 285
 Bradford, Yvonne 430*
 Bradshaw, Richard 326
 Brinkley, James F. 444
 Brinkman, Ryan R. 240, 316, 377
 Brochhausen, Mathias 167, 183
 Brush, Matthew H. 101, 234*
 Bult, Carol J. 285
 Bulu, Hakan 159*

 Canada, Brian 263, 288*, 441*
 Cappadona, Nick 296
 Caruso, Brian 296
 Castro, Alexander Garcia 399
 Caudy, Michael 285
 Ceusters, Werner 71, 252, 309
 Chalmers, Robert J.G. 133
 Chapman, Chris 258

 Charlet, Jean 241, 276
 Chatr-Aryamontri, Andrew 263
 Chaudhri, Vinay 273
 Chen, Shanshan 260
 Cheng, Keith C. 263, 288, 441
 Cheng, Shunfeng 340*
 Chibucos, Marcus 237*
 Choquet, Rémy 241
 Chung, Caty 209
 Chute, Christopher G. 133, 266, 292, 329, 332
 Clark, Kimberly 332
 Conn, P. Michael 263
 Cook, Daniel L. 41*
 Cooper, Timothy 441
 Corday, Karen 101
 Cormont, Sylvie 276
 Corson-Rikert, Jon 260, 296
 Courtot, Mélanie 240*, 316*, 377*
 Couto, Francisco M. 3, 117, 290, 300
 Cowell, Lindsay G. 255, 387, 393
 Cross, Valerie 125*, 290
 Cruz, Isabel F. 125, 290*

 D'Eustachio, Peter 285
 Dahdul, Wasila 230, 428*
 Das, Diganta 340
 Davis, Allan Peter 231
 de Bono, Bernard 393
 de Matos, Paula 249
 Deans, Andrew R. 422
 Dejori, Mathaeus 364
 Dekker, Adriano 249
 Dhamanaskar, Alok 246
 Dhombres, Ferdinand 241*, 276
 Diehl, Alexander D. 11, 25, 259, 285, 370
 Ding, Ying 260
 Doelling, Erin D. 258
 Doi, Koji 79
 Dolan, Mary E. 231
 Dolinski, Kara 263

 Dorf, Michael 292, 302
 Drabkin, Harold J. 285
 Duke, Jon D. 329
 Dunn, Kim 436

 eagle-i Consortium 260
 Ellisman, Mark 263
 Emilsson, Pia 324*
 Ennis, Marcus 249
 Eppig, Janan T. 231, 263
 Essaid, Shahim 244*
 Evsikov, Alexei 285

 Faria, Daniel 300
 Ferguson, Ray 292, 302
 Ferreira, João D. 117*, 300
 Forsberg, Kerstin 324
 Frank, Robert 55
 Frishkoff, Gwen 55*

 Galdzicki, Michal 41
 Garcia, Alexander 87
 Garcia, Jael 87
 Garimalla, Swetha 147, 252
 Gennari, John H. 41
 German, J. Bruce 446
 Giglio, Michelle 237
 Giraldo, Olga 87
 Gkoutos, Georgios V. 79
 Goldenkrantz, Andrew 273
 Goldman, Julian M. 335
 Gomadam, Karthik 367
 Gonzalez, Graciela 447
 Grego, Tiago 300
 Grenon, Pierre 393
 Grethe, Jeffrey S. 263, 349
 Gross, Anika 109*
 Gupta, Amarnath 349
 Guttula, Chaitanya 246*
 Haendel, Melissa 101, 234, 260, 263, 370
 Hammond, Maya 258

- Hanauer, Marc 241
Hancock, John 416
Harb, Omar S. 95
Hartung, Michael 109
Hastings, Janna 71*, 201*, 249*, 357, 399
Haug, Kenneth 249
Hayamizu, Terry F. 411*
He, Yongqun 25, 33, 279, 304, 309*
Hoehndorf, Robert 79, 183
Hogan, William R. 147*, 252*
Holland, Peter W.H. 416
Horridge, Matthew 294
Hu, James 237
Hu, Xueheng 125
Huang, Hongzhan 285
- Iadonato, Shawn 263
Iannone, Luigi 294
Imai, Takeshi 63
Imam, Fahim T. 349*
Imms, Ryan 47
Ireland, Amelia 370
- Jain, Prateek 367
Jiang, Guoqian 133*, 329*, 332
Johnson, Tenille 101, 234
Jonquet, Clement 302
Josephs, Zara 249
Jupp, Simon 294*
- Kaput, Jim 446
Karlsson, Daniel 153*
Kemnitz, Joseph 263
Kim, Jee-Hyub 357
Kirsten, Toralf 109
Kissinger, Jessica C. 246
Klein, Julie 294
Kobayashi, Norio 79
Koleti, Amar 209
Kou, Hiroko 63
Kozaki, Kouji 63, 79
Kremers, Hilal M. 266
Kushida, Tatsuya 79
- Kutz, Oliver 399*
- Landgrebe, Jobst 139*
Lange, Matthew 446*
Lapp, Hilmar 230, 426, 428
Larson, Stephen D. 263, 349
LaSalle, Bernard 326*
Lepage, Eric 276
LePendu, Paea 55, 435*
Lemmon, Vance 209
Levin, Mikhail K. 255*, 387
Lin, Yu 33*, 258*, 304, 440*
Lowe, Brian 260, 296*
Luciano, Joanne 191
Mabee, Paula M. 230, 428
Madani, Sina 436*
Mader, Christopher 209
Magness, Charles 263
Malone, James 19, 25, 416
Manzoor, Shahid 370
Martone, Maryann E. 263, 349
Masci, Anna Maria 255, 387*, 393*
Masuya, Hiroshi 79*
Mattingly, Carolyn 231
McCusker, James P. 191*
McGuinness, Deborah L. 191
Meehan, Terrence F. 11*, 25, 259*, 370
Mejino Jr., Jose L.V. 41, 444
Midford, Peter E. 230, 426, 428
Mikó, István 422*
Miller, John A. 246
Mirhaji, Parsa 436
Mitchell, Stella 260*, 296
Mizoguchi, Riichiro 63*, 79
Mosskowski, Till 399
Mulligan, Kevin 71
Mungall, Christopher J. 11, 259, 263*, 279, 370*, 377
Murphy, Shawn P. 266
Musen, Mark A. 292, 298, 302, 435
- Nair, Jithun 298*
Natale, Darren A. 285
- Neal, Maxwell L. 41
Neuhaus, Fabian 201
Nichols, B. Nolan 444*
Noy, Natalya 292, 298
- Ogbuji, Chimezie 217*
Ohe, Kazuhiko 63
Okuda, Yoshihiro 79
Olry, Annie 241
Osumi-Sutherland, David 370, 412*, 425*
Overton, James 438*
Owen, Gareth 249
Owen, Stuart 294
- Pacheco, John 273
Palla, Ravi 364*
Panahiazar, Maryam 125
Pang, Chao 19*, 25, 416
Parikh, Priti 125
Parkinson, Helen 19, 25, 416
Pathak, Jyotishman 266*, 329
Pecht, Michael 340
Pesquita, Catia 3*, 290, 300
Petrova, Mila 433*
Proietti, Anna Barbara de Freitas Carneiro 227
Przydzial, Magdalena 209
Puri, Colin 367*
- Rahm, Erhard 109
Rath, Ana 241
Rebholz-Schuhmann, Dietrich 357
Rector, Alan L. 133
Rejack, Nicholas 260
Richardson, Joel 231
Ringwald, Martin 411
Roach, Jeffrey 393
Roberts, Natalia V. 285
Rocca, Walter A. 266
Ruttenberg, Alan 255, 279, 285, 316, 377, 381, 387
- Sahoo, Satya S. 269*

Sakurai, Kunie 209	Storey Margaret-Anne 292	Wakana, Shigeharu 79
Sarntivijai, Sirarat 25*, 309	Stroe, Cosmin 125, 290	Waki, Kayo 63, 79
Schanstra, Joost 294	Summers, Ron 47	Wang, James Z. 288
Schleich, Jean-Marc 47		Wang, Rui 246
Schleicher, John 288	Tanaka, Nobuhiko 79	Weininger, Sandy 335
Schneider, Luc 167*	Tao, Cui 332*	Westerfield, Monte 230, 263, 428
Schofield, Paul 416	Tariq, Shariq A. 147, 252	Whetzel, Patricia L. 292*, 298, 302*
Schulz, Stefan 153, 183*, 357	Tavares, Bruno 300*	
Schürer, Stephan 25, 209	Tecuci, Dan 364	Whitehead, Susan F. 335
Segerdell, Erik 101	Thomas, Georgia 288, 441	Wieggers, Thomas C. 231
Seltmann, Katja C. 422	Toldo, Luca 309	Willaert, Brian N. 266
Seyed, A. Patrice 175*, 273*	Torniai, Carlo 101*, 234, 260, 263, 370	Wilson, Melanie 101, 260
Shaffer, Chris 101, 234	Toyoda, Tetsuro 79	Wolstencroft, Katy 294
Shah, Nigam H. 292, 302, 435	Travillian, Ravensara 416*	Wong, David Tai Wai 381
Shapiro, Stuart C. 175	Troyanskaya, Olga 263	Worthington, Miles 296
Sharma, Deepak 332	Tudorache, Tania 298	Wu, Cathy H. 285
Shet, Vinay 364	Turner, Judith 263	
Sheth, Amit 125	Turner, Steve 249	Xiang, Zuoshuang 25, 33, 279*, 304*, 309
Siegele, Deborah 237	Tyers, Mike 263	
Silwal, Pramit 125		Yamagata, Yuki 63
Sittig, Dean F. 436	Uetz, Peter 237	Yan, Ying 357*
Smith, Barry 71, 139, 227, 285, 292, 381	Urbero, Bruno 241	Yao, Robert 447*
Smith, Cynthia 231		Yawn, Barbara P. 266
Smith, Robin 209	Van Slyke, Ceri 430	Yeh, Peter Z. 367
Sojic, Aleksandra 399	Vandenbussche, Pierre-Yves 241, 276*	Yoder, Matthew J. 422
Solbrig, Harold R. 133		Youn, Cherie 302
Spackman, Kent 133	Vasilevsky, Nicole 101, 234	
Steinbeck, Christoph 201, 249, 357	Vempati, Uma 25, 209*	Zhang, Jian 285
Stevens, Robert 294	Verma, Kunal 367	Zheng, Jie 95*, 246
Stoeckert Jr., Christian J. 95, 246	Vision, Todd J. 230, 428	Zweifel, Adrienne 237
	Visser, Ubbo 209	
	VIVO Collaboration 260, 296	

* First Author

