Semantic Trilogy '13

Proceedings

International Conference on Biomedical Ontology 2013

Workshops

Montreal, Quebec, Canada July 6th—12th 2013

UNIVERSITÉ Concordia UNIVERSITY

Semantic Trilogy '13

Definitions in Ontologies

Montreal, Quebec, Canada July 6th—12th 2013

UNIVERSITÉ Concordia UNIVERSITY

INTERNATIONAL WORKSHOP ON DEFINITIONS IN ONTOLOGIES (DO 2013)

Proceedings

edited by

Selja Seppälä
and
Alan Ruttenberg

DO 2013 is held in conjunction with the
4th International Conference on Biomedical Ontology (ICBO 2013)

Concordia University
Montreal, Quebec, Canada

July 7, 2013

FNSNF
Swiss National Science Foundation

University at Buffalo
The State University of New York

Preface

Ontologies built using OBO Foundry principles are advised to include both formal (logical) definitions, as well as natural language definitions. Depending on the effort, one or the other can be underrepresented. Possible explanations to this bottleneck are the high cost of producing well-written definitions; an insufficient understanding of the nature of natural language definitions or of logic; the lack of an operational theory of definitions; the lack of studies that evaluate usability and effectiveness of definitions in ontologies; a paucity of tools to help with definition authoring and checking. Producing natural language definitions is time-consuming, costly and prone to all kinds of inconsistencies. Producing logical definitions that are effective, correct, and communicative is also difficult. It is therefore worth exploring different ways of assisting, with automation, creation and quality control of definitions.

This workshop gathers interested researchers and developers to reflect upon general themes as the selection and modeling of defining information; the relation between definitions in specific domains as opposed to domain-independent definitions; the theoretical underpinnings of definitions; tools that can facilitate relating logical and natural language definitions. In addition, we wanted to encourage participation by different communities using definitions so that their needs and solutions can be exposed. This interdisciplinary goal proved successful, as each of the three selected papers represents a different community: ontology, terminology, and natural language processing (NLP).

Topics

Topics of interest were split between foundational aspects, pragmatic issues and user perspectives. Below we list some possible topics.

Foundational aspect

- Theories of definition and their implications for the defining practice
- Realist versus conceptualist approaches in definition writing
- Definition modeling: what kinds of information are defining
- Domain-independent versus domain-specific definition models
- Formal versus natural language definitions

Pragmatic issues

- Quality control in definitions
- Ways of evaluating definitions
- Comparison and evaluation of different definition production techniques: handwritten, automatically generated from formal definitions, extracted from corpora or constructed from information retrieved from corpora
- Methods and tools to automate definition production and checking
- (Multilingual) definition generation

- Information retrieval for definition production
- Use of definition models to facilitate information retrieval
- Definition extraction from corpora
- Interactions between ontologies and lexical resources (WordNet, FrameNet)
- Consequences/Strategies of giving necessary versus necessary and sufficient definitions, or simply sufficient definitions
- Coordination of logical and textual definitions
- Alternatives to and variants of definitions: elucidations, explanations, glosses, figures

User perspectives

- Assessment of definitions used in current practice
- Balancing needs of within discipline use and wider use of definitions
- Use of specialized terminology versus general vocabulary
- Presentation of definitions to different user audiences
- Alternatives/Augmentations of textual definitions, such as figures and images for anatomy, where textual definitions may be harder to formulate

The articles published in these proceedings cover three distinct aspects of definitions in ontologies, and include a survey report summary on defining practices in the ontology community conducted in preparation of the workshop:

Summary of the survey results:

Selja Seppälä and Alan Ruttenberg

Survey on Defining Practices in Ontologies: Report Summary

Automatically populating ontologies with multilingual definitions:

Guoqian Jiang, Harold Solbrig and Christopher Chute

A Semantic Web-Based Approach for Harvesting Multilingual Textual Definitions from Wikipedia to Support ICD-11 Revision

Displaying user-oriented defining contents using domain-specific templates:

Antonio San Martín and Pilar León Araúz

Flexible Terminological Definitions and Conceptual Frames

Enhancing term retrieval through the definition's contents:

Gerardo Sierra-Martinez and Laura Elena Hernandez-Dominguez

Automatic Construction of the Knowledge Base of an Onomasiological Dictionary

In addition to these papers, the workshop's program includes three presentations aimed at widening the scope of the issues to be discussed:

Pragmatic aspects of defining diseases in the Disease Ontology:
Lynn M. Schriml
Definition of Disease Terms

Domain- and language-independent definition templates:
Selja Seppälä
Using BFO Categories for Creating Generic Definition Templates

Definition authoring tools:
Alan Ruttenberg
Tool Support for Definition Authorship and Management: Desiderata

Our invited speaker covers some foundational aspects of definitions in ontologies:
Barry Smith
Introduction to the Logics of Definitions

We expect to have a rich and insightful discussion with the participants and the attendees of the workshop that will yield a prioritized list of needs and useful recommendations regarding the defining practices in ontologies.

Acknowledgements

This workshop would not have been possible without the thorough reviews of the 18 scholars in the domains of ontology, natural language processing, terminology and lexicography that constituted the program committee. Three referees reviewed each of the four submissions we received. We would like to thank them for their participation and support to this workshop. We would also like to thank the ICBO organizers for hosting the workshop and providing the logistic assistance, as well as the Swiss National Science Foundation and the State University of New York at Buffalo for their support.

July 2013

Selja Seppälä
Alan Ruttenberg

Organizing Committee

Selja Seppälä (University at Buffalo, USA)
Alan Ruttenberg (University at Buffalo, USA)

Program Committee

César Aguilar (Pontificia Universidad Católica de Chile)
Nathalie Aussenac-Gilles (National Center for Scientific Research (CNRS), France)
Caroline Barrière (CRIM, Canada)
Thomas Bittner (University at Buffalo, USA)
Mélanie Courtot (British Columbia Cancer Research Centre, Canada)
Christiane Fellbaum (Princeton University, USA)
Natalia Grabar (Université de Lille 3, France)
Janna Hastings (European Bioinformatics Institute, Cambridge, UK)
Marie-Claude L'Homme (Université de Montréal, Canada)
James Malone (European Bioinformatics Institute, Cambridge, UK)
Alexis Nasr (Aix Marseille Université, France)
Fabian Neuhaus (National Institute of Standards and Technology (NIST), USA)
James Overton (Knocean, Toronto, Canada)
Richard Power (The Open University, UK)
Patrice Seyed (Tetherless World Constellation, Rensselaer Polytechnic Institute, USA)
Robert Stevens (The University of Manchester, UK)
Allan Third (The Open University, UK)
Sandra Williams (The Open University, UK)

DO 2013 PROGRAM

July 7, 2013

8:30–8:45 Welcome

8:45–9:00 Lynn M. Schriml
Definition of Disease Terms

9:00–9:15 Gerardo Sierra-Martinez and Laura Elena Hernandez-Dominguez
Automatic Construction of the Knowledge Base of an Onomasiological Dictionary

9:15–9:30 Guoqian Jiang, Harold Solbrig and Christopher Chute
A Semantic Web-Based Approach for Harvesting Multilingual Textual Definitions from Wikipedia to Support ICD-11 Revision

9:30–9:45 Antonio San Martín and Pilar León Araúz
Flexible Terminological Definitions and Conceptual Frames

9:45–10:15 **Discussion**

10:15–10:45 Coffee Break

10:45–11:15 **Invited Speaker**: Barry Smith
Introduction to the Logic of Definitions

11:15–11:30 Selja Seppälä
Using BFO Categories for Creating Generic Definition Templates

11:30–11:45 Alan Ruttenberg
Tool Support for Definition Authorship and Management: Desiderata

11:45–12:00 **Discussion**

Survey on Defining Practices in Ontologies

– Report Summary –

This document reports the results of a survey on defining practices in ontologies conducted in preparation of the International Workshop on Definitions in Ontologies (DO 2013) held on July 7, 2013, in Montreal, in conjunction with the Fourth International Conference on Biomedical Ontologies 2013 (ICBO2013), itself part of the Semantic Trilogy '13 event.

1. Background

Ontologies built using OBO Foundry principles are advised to include both formal (logical) definitions and natural language definitions. Depending on the effort, one or the other can be underrepresented. Possible explanations to this bottleneck are the high cost of producing well-written definitions; an insufficient understanding of the nature of natural language definitions or of logic; the lack of an operational theory of definitions; the lack of studies that evaluate usability and effectiveness of definitions in ontologies; a paucity of tools to help with definition authoring and checking.

Producing natural language definitions is time-consuming, costly and prone to all kinds of inconsistencies. Producing logical definitions that are effective, correct, and communicative is also difficult. It is therefore worth exploring different ways of assisting, with automation, creation and quality control of definitions.

Accordingly, we thought it would be useful to gather interested researchers and developers to reflect upon general themes as the selection and modeling of defining information; the relation between definitions in specific domains as opposed to domain-independent definitions; the theoretical underpinnings of definitions; tools that can facilitate relating logical and natural language definitions. In addition, we wanted to encourage participation by different communities using definitions so that their needs can be exposed.

To address these issues, we organized a half-day workshop aimed at discussing questions, ideas and existing projects regarding definitions in ontologies. The expected outcomes of the workshop were to get an overall view of the needs of the users so as to best orient research on the definition authoring side, as well as to get a diagnosis of the difficulties faced by the latter in order to guide groundwork on definitions and their production.

We present here the results of the survey on defining practices that was conducted in preparation of the general discussion at the workshop.

2. Objectives

The objective was to gather information on the practices and needs of the ontology community with respect to definitions – logical and textual – in order to guide the discussion session aimed at creating a prioritized list of needs and best practices in definitions. We invited the ontology community to give us feedback on their experience by filling in a questionnaire published on the Internet. The web-based survey was sent to several ontology lists; 14 people responded to the questionnaire. The small number of participants does not allow us to draw statistically

significant conclusions; their answers are nevertheless indicative of the practices and needs related to definitions in ontologies.

3. Methodology

The 15 questions – some of which include sub-questions – of the questionnaire can be grouped into three larger categories:

- **User-oriented questions:** types of users; their role in the ontology project on which they are working; their use of logical and/or textual definitions; their training in logical and/or textual definition authoring; the kind of assistance needed with respect to definitions in ontologies. [Q: 1, 2, 8a, 9a-d, 10a-b, 11a-b, 14]
- **Ontology-oriented questions:** inclusion of logical and/or textual definitions in the ontologies. [Q: 3, 4a-c, 5a-b]
- **Definition-oriented questions:** usefulness of logical and textual definitions; major problems in definitions; desired enhancements to textual definitions. [Q: 6, 7, 8b, 12, 13]

Several types of questions were asked: closed yes/no questions; multiple choice questions with single or multiple answers, and open-ended textual (qualitative) questions.

4. Summary of the results

4.1. Users and their needs

4.1.1. Respondents' profile (Q1-2)

Most of the respondents work as ontologists regardless of their primary profession. They are thus more likely to be involved in definition authoring and to express needs related to these activities, which is confirmed by the results to the other questions in this section.

4.1.2. Use of definitions (Q8a-b)

The majority of the respondents report using – consulting or writing – definitions 'often', which is indicative of the fact that definitions are central to the ontology work.

Two types of uses seem to emerge: mostly internal uses related to the activity of ontology development, and, to a lesser extent, external uses related to the application of ontologies.

The answers suggest that respondents are primarily concerned with logical definitions. The lesser use of textual definitions may be due to their lacking quality. These results suggest, in turn, that the roles of the term, the logical definition and the textual one in ontologies could be more precisely defined.

4.1.3. Definition consultation (Q9a-d)

All of the respondents report using logical and/or textual definitions to get a clear understanding of the terms in the ontologies; moreover, the majority of them report using definitions 'often' rather than 'sometimes'.

The use of logical definitions is quantitatively closer to 'very often' than to 'rarely' (7/12 vs. 5/12 respondents). However, the use of textual definitions is even more frequent than that of logical definitions (9/11 respondents).

The frequent use of both logical and textual definitions seems to indicate that they play an important role in the proper understanding of what is represented in the ontologies.

4.1.4. Definition writing (Q10a-b)

The majority of the respondents report engaging in definition authoring activities.

The defining activity is not only limited to definition creation, generally, from texts and consultation of experts; it also includes definition revision and 'translation' of textual definitions to/from logical ones.

As for the defining form, the classical definition structure – genus + differentia – is the preferred one.

4.1.5. Training in definition writing (Q11a-b)

Half of the respondents have had no training in definition writing. Among the other half of the respondents, most have had training in both logical and textual definition writing, and one only in logical definition writing. In only a few cases the training in definition writing was ontology-oriented. It would thus be interesting to create this kind of specific training.

4.1.6. Users' needs (Q14)

The ontology community would mostly welcome general principles for definition writing. Half of the respondents were also interested in ontology-specific training for writing logical definitions. The results also suggest that training and tools related to textual definitions tend to be considered as nice-to-have but not as important as assistance with logical definitions.

4.2. Ontologies and Definitions

4.2.1. Kinds of ontologies (Q3)

Most of the respondents work on ontologies related to the biomedical domain; two work on an upper level ontology, the Basic Formal Ontology. The other ontologies cover varied areas.

4.2.2. Importance of definitions in ontologies

4.2.2.1. Importance of logical definitions (Q4a-c)

The majority of the ontologies on which the respondents answered these questions include logical definitions. However, in most of the ontologies, less than half of the entities are logically defined; only in one are 75-100% of the entities defined.

These results suggest to us that more logical definitions will be added in the future, in particular if the ontology developers want to comply with the OBO Foundry principles. Hence, authoring tools that allow for the semi-automatic creation of logical definitions would probably be helpful.

4.2.2.2. Importance of textual definitions (Q5a-b)

All the ontologies on which the respondents answered these questions except one have textual definitions. Moreover, by contrast with logical definitions, the textual definitions are well represented: in 10/12 ontologies, more than half of the entities are defined with a textual definition; the coverage rate in 2/3 of all the ontologies is even comprised between 75% and 100% of the entities.

These results tend to indicate that the needs related to textual definitions may be less pronounced than those related to logical definitions.

4.3. Definitions in Ontologies

4.3.1. Usefulness of definitions (Q6-7)

Both logical and textual definitions are subjectively considered by the respondents as extremely important in ontologies.

4.3.2. Problems with definitions (Q12)

Four large types of problems were mentioned by the respondents. These are related to:

1. the information content of textual definitions
   - insufficiently informative
   - too informative/too complex
   - outdated
   - absence of standard defining patterns
2. logical issues
   - vague
   - circular
   - self-contradictory
3. the writing and style of the definitions
   - poorly written
   - inconsistent in style
4. coverage
   - multiple definitions
   - absence of definitions

4.3.3. Desired enhancements in textual definitions (Q13)

The most frequently mentioned desired enhancements to textual definitions relate (i) to their authoring methods – the creation of definition templates –, and (ii) to their content and form – an increase in the readability of the definitions. The latter enhancement includes not only stylistic matters, but also adaptability of the defining vocabulary to different types of users, which is also related to the adaptability of the defining content. Current user-oriented trends of research in terminology and lexicography could be helpful in this respect.

Among the other mentioned enhancements, we note the development of tools or methods to convert textual definitions to/from logical ones, issues that have started to be explored in the ontology community. Finally, the inclusion of examples is also mentioned, although this enhancement is not as such related to definitions; it may however be indicative of definitions

that are not explicit and content-wise not rich enough to be useful to the users – although in some (or maybe many) cases it might not be related to the lacking of definitions at all, only to the fact that examples tend to fulfill a different cognitive need.

4.3.4. Further comments and suggestions (Q15)

Tools should be developed to help ontology developers implement general principles on definitions.

5. Conclusion

In conclusion, this survey on defining practices in ontologies suggests that definitions are central to ontologies, not only for computational reasons, but also for their proper development and use by humans.

Concerning users' needs, the survey results indicate that it would be valuable to establish ontology-oriented defining principles and manuals, backed up with tools to support ontology developers in implementing the recommendations. Moreover, specific ontology-oriented definition writing training courses or tutorials would also be among the priorities, in particular for logical definitions.

Finally, the current rather low definition coverage rate in ontologies suggests to us that, in light of the standards of good practice in ontology development, more logical, but also textual, definitions will (or, at least, should) be added in the future. Therefore, research efforts could be geared towards developing tools that allow for the (semi-)automatic creation of definitions, for example by generating textual definitions to/from logical ones.

A detailed analysis of the results of the questionnaire is available on the DO 2013 workshop's website: http://definitionsinontologies.weebly.com.

A Semantic Web-Based Approach for Harvesting Multilingual Textual Definitions from Wikipedia to Support ICD-11 Revision

Guoqian Jiang[1,*] Harold R. Solbrig[1] and Christopher G. Chute[1]
[1] Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN

ABSTRACT

In the beta phase of the 11th revision of International Classification of Diseases (ICD-11), the World Health Organization (WHO) intends to accept public input through a distributed model of authoring, in which creating textual definitions for ICD categories is a core use case. In a previous study, Wikipedia has been demonstrated as a useful source for textual definitions of diseases. The objective of the study is to develop and evaluate a semantic web-based approach for harvesting multilingual textual definitions from Wikipedia to support ICD-11 revision and its public review. In a prototype implementation, we developed a semantic web service application known as LexReview that automates the harvesting process in a dynamic way through invoking and integrating three online web services: 1) WHO ICD-11 content services; 2) NCBO BioPortal annotation services; and 3) DBpedia SPARQL endpoint query services. The Simple Knowledge Organization System (SKOS) lexical and mapping properties are used to represent the harvested definitions. The LexReview service application could be extended to integrate the textual definitions from other resources and subsequently consumed by a review application to support ICD-11 revision.

1 INTRODUCTION

The 11th revision of International Classification of Diseases (ICD-11) was officially launched by the World Health Organization (WHO) in March 2007 (1). The beta phase of the ICD-11 revision started in May 2012, and WHO intends to accept public input through a distributed model of authoring. An ICD-11 Beta Browser application has been developed and released by WHO (2). The browser provides simple commenting functionality to allow the domain professionals to make comments on existing contents, and it intends to introduce more social computing capabilities.

Lexical properties of ICD categories including titles, synonyms, and textual definitions should be reviewed following a standard and homogeneous terminological approach. The provision of textual definitions has been regarded as one of important criteria for measuring the quality of a terminology/ontology (3). In our previous study (4), we demonstrated that the textual definitions from the Unified Medical Language System (UMLS) (5), the formal definitions of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (6) and the linked open data (LOD) from DBpedia (7) are potentially useful resources for supporting ICD-11 textual definitions authoring. We argued that the ICD-11 project might potentially take advantage of the crowd-souring model of Wikipedia (8). Using this model, each ICD-11 category would be seeded as a Wikipedia page for public input and the definitions of ICD categories would be harvested using the DBpedia.

The objective of the study is to develop and evaluate a semantic web-based approach for harvesting multilingual textual definitions from Wikipedia to support ICD-11 revision and its public review. In a prototype implementation, we developed a semantic web service application known as LexReview that automates the harvesting process in a dynamic way through invoking and integrating a number of online web services: 1) WHO ICD-11 content services; 2) NCBO BioPortal annotation services; and 3) DBPedia SPARQL endpoint query services. The Simple Knowledge Organization System (SKOS) lexical and mapping properties are used to represent the harvested definitions.

2 BACKGROUND

2.1 WHO ICD-11 Content Model and Services

An ICD-11 content model has been developed by WHO to present the knowledge that underlies the definitions of an ICD entity. The content model is composed of three layers: a foundation component, a linearization component and an ontological component (9). The foundation component stores the full range of knowledge of all classification units in ICD. The linearization component corresponds to the classical print versions of ICD. The ontological component provides references to formal definition of terms and relationships. Currently, there are 13 defined main parameters in the content model to describe a category in ICD, in which "Textual Definitions" is one of main parameters for describing an ICD category.

Recently, an ICD URI scheme is proposed for naming and supporting web services by WHO. A base URI of http://id.who.int has been proposed, with http://id.who.int/icd/schema as the prefix for the vocabulary terms that related to ICD classification efforts maintained by WHO, http://id.who.int/icd/entity for the fundamental foundation entities related to ICD concepts.

2.2 BioPortal Annotation Services

The National Center for Biomedical Ontology Annotator is an ontology-based web service for annotating the textual biomedical data with biomedical ontology concepts (10, 11).

* To whom correspondence should be addressed: jiang.guoqian@mayo.edu

The biomedical community can use the Annotator service to tag datasets automatically with concepts from more than 300 ontologies coming from the two most important biomedical ontology & terminology repositories: the Unified Medical Language System (UMLS) Metathesaurus and NCBO BioPortal. Such annotations contribute to create a biomedical semantic web that facilitates translational scientific discoveries by integrating annotated data. In this study, the Medical Subject Headings (MeSH) (12) was configured to annotate the preferred labels of ICD-11 categories.

2.3 DBpedia SPARQL Endpoint

DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web (7). DBpedia adopts Semantic Web Linked Open Data technology and its datasets are rendered in RDF format and can be accessed online via a public SPARQL query endpoint at http://dbpedia.org/sparql. The endpoint is provided using OpenLink Virtuoso as the back-end RDF database engine.

DBpedia also defines an ontology to organize its datasets. The ontology is a shallow, cross-domain ontology and covers 359 classes that form a subsumption hierarchy and are described by 1,775 different properties. In this study, we used one of the classes http://dbpedia.org/ontology/Disease and extracted all instances of the class for obtaining textual definitions.

2.4 Semantic Web Technologies

The World Wide Web consortium (W3C) is the main standards body for the World Wide Web (13). The goal of the W3C is to develop interoperable technologies and tools as well as specifications and guidelines to lead the web to its full potential. The resource description framework (RDF), web ontology language (OWL), and SPARQL (a recursive acronym for **S**PARQL **P**rotocol and **R**DF **Q**uery **L**anguage) specifications have all achieved the level of W3C recommendations, and are becoming generally accepted and widely used.

The SKOS data model views a knowledge organization system as a concept scheme comprising a set of concepts (14).The vocabulary used in the SKOS data model is a set of URIs that specifies the notion of SKOS concepts, concept schemes, lexical labels, notations, documentation properties and semantic relations. SKOS data are expressed as RDF triples. An increasing number of SKOS datasets in RDF are publicly available.

3 SYSTEM ARCHITECTURE

Figure 1 shows the system architecture of our approach. The LexReview service appplication invoked and integrated mainly three web services: 1) WHO ICD-11 content services for retrieving preferred label and definition for a target ICD entity; 2) NCBO BioPortal annotation services for retrieving the MeSH term annotation and its ID; and 3) DBpedia SPARQL endpoint query services for retrieving textual definitions by MeSH ID.

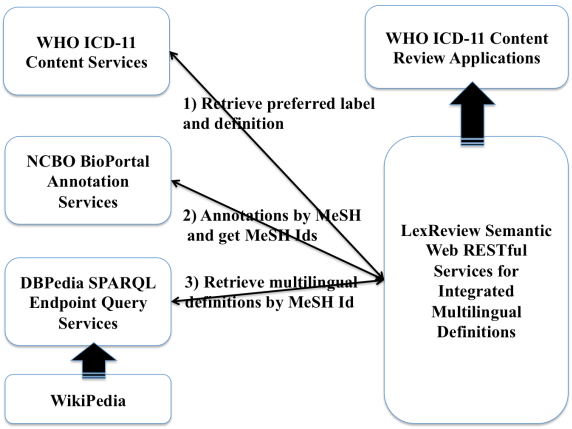

Figure 1. System architecture of the LexReview service application.

4 PROTOTYPE IMPLEMENTATION

The LexReview service application was implemented using a Java-based RESTful web services JAX-RS API known as Jersey (15) and a Jena ARQ API (16) that is a Java-based query engine for Jena that supports SPARQL RDF query language.

The service application accepts a standard URI of a single ICD entity as input. For example, the URI - http://id.who.int/icd/entity/718946808 represents an ICD entity Angina pectoris. Figure 2 shows the HTML rendering of the ICD entity Angina pectoris dispalyed through a web browser.

The content of an ICD-11 entity can be accessed through Content Negotiation that is a mechanism of RESTful services that makes it possible to serve different representation of a resource at the same URI. The WHO ICD content services provide the content representation in the formats of HTML, RDF and JSON. First, the system retrieved the title and definition of a target ICD entity based on its RDF rendering, in which the SKOS lexical properties skos:prefLabel and skos:definition are used to represent the values.

Second, the system invoked NCBO BioPortal annotation services using the title of a target ICD entity as the input. The annotation services were configured to use the ontology MeSH only and the semantic types within the semantic group Disorders (17)(see Table 1). The annotation services provide a score for each annotation that is the weight based on the annotation context. In this prototype implmentation, we harvested those annotation with the score=10, meaning that a direct annotation is matched with a concept preferred

name. We then retrieved the MeSH ID, preferred name and URI of each annotation.

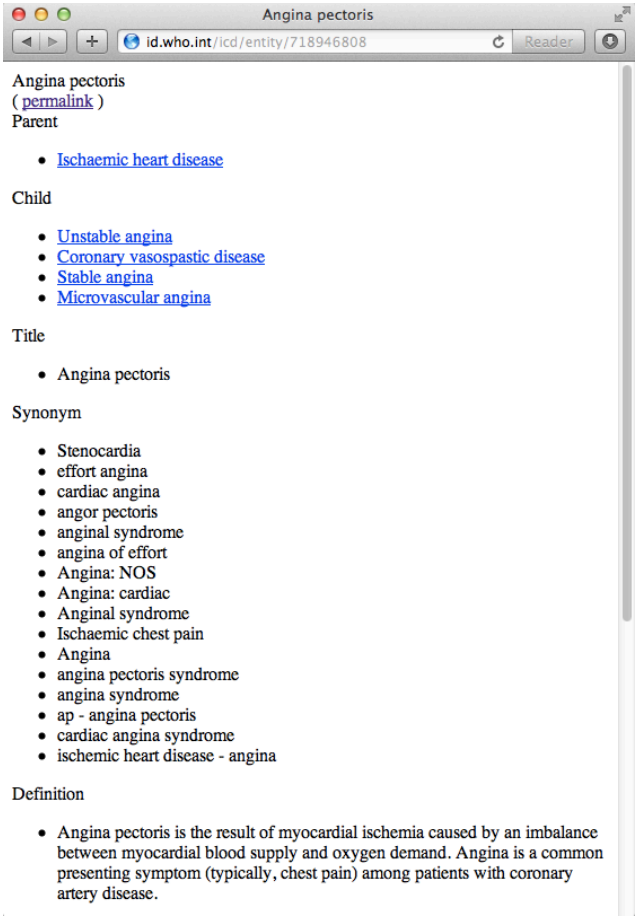

Figure 2. The HTML rendering of the ICD entity Angina pectoris.

Table 1. A list of semantic types within the semantic group Disorders

DISO|Disorders|T020|Acquired Abnormality
DISO|Disorders|T190|Anatomical Abnormality
DISO|Disorders|T049|Cell or Molecular Dysfunction
DISO|Disorders|T019|Congenital Abnormality
DISO|Disorders|T047|Disease or Syndrome
DISO|Disorders|T050|Experimental Model of Disease
DISO|Disorders|T033|Finding
DISO|Disorders|T037|Injury or Poisoning
DISO|Disorders|T048|Mental or Behavioral Dysfunction
DISO|Disorders|T191|Neoplastic Process
DISO|Disorders|T046|Pathologic Function
DISO|Disorders|T184|Sign or Symptom

Third, when the system had a MeSH term annotated for a target ICD entity, the system invoked the DBpedia SPARQL enpoint to retrieve the textual definitions of a DBpedia entry coded in a MeSH ID. Figure 3 shows the SPARQL query used to retrieve mulilingual textual definitions from the instance entries of a DBpedia class "Disease" (i.e., http://dbpedia.org/ontology/Disease).

```
PREFIX dbpedia: <http://dbpedia.org/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?dbpediaEntry ?abstract ?wikipediaPage WHERE {
  ?dbpediaEntry a <http://dbpedia.org/ontology/Disease> .
  ?dbpediaEntry dbpedia:abstract ?abstract .
  ?dbpediaEntry foaf:isPrimaryTopicOf ?wikipediaPage .
  ?dbpediaEntry dbpedia:meshId "D018805"@en .
  FILTER (lang(?abstract)="ar" || lang(?abstract)="zh" || lang(?abstract)="en"
        ||lang(?abstract)="fr" ||lang(?abstract)="ru" ||lang(?abstract)="es")
}
```

Figure 3. The SPARQL query used to retrieve multilingual textual definitions from DBpedia for a MeSH ID (e.g., D018805 for the MeSH term Sepsis)

Here, we asserted that the values of the predicate dbpedia:abstract are candidates for textual definitions. We used the language tags as a filter to retrieve those textual definitions in six official languages adopted by the WHO (18), i.e. "ar" standing for Arabic, "zh" for Chinese, "en" for English, "fr" for French, "ru" for Russian, and "es" for Spanish.

Finally, we represented the MeSH mapping based on BioPortal annotation services and the multilingual textual definitions retrieved for a target ICD category in RDF format, in which the SKOS lexical and mapping properties (skos:prefLabel, skos:definition, skos:closeMatch, skos:exactMatch) are used. We then exposed the RDF rendering through a RESTful service API. Figure 4 shows an example of RDF rendering of multilingual textual definitions for a target ICD-11 entity Angina pectoris. As illustrated in the figure, we used the predicate skos:closeMatch to represent the relationship between the target ICD entity and its MeSH annotation http://purl.bioontology.org/ontology/MSH/D000787. We used the predicate skos:exactMatch to represent the relationship between the MeSH annotation with the DBpedia entry http://dbpedia.org/resource/Angina_pectoris because they share the same MeSH ID. There are 11 definition entries in 5 languages available for the DBpedia entry and the predicate skos:definition is used to represent them. In addition, we also put the original title and definition of the target ICD entity in the RDF rendering using the predicates skos:prefLabel and skos:definition.

The prototype implementation will be accessible soon through http://informatics.mayo.edu/rest/project/icd11/lexreview/definition?uri=http://id.who.int/icd/entity/718946808, in which the uri parameter can be replaced by any other ICD entity URIs.

Table 2 shows a list of ICD-11 entity examples (n=10) that have Wikipedia definition matches. The first column in Table 2 shows the ICD-11 entity URI and its preferred label; the second column shows the corresponding Wikipedia URI for each ICD-11 entity matched by the system, and the codes for available languages; the third column shows the MeSH ID being an anchor between an ICD-11 entity and an Wikipedia entry. For each ICD-11 entity in Table 2, the Wikipedia definition entries are available at least in two language codes (range from 2-5 codes). The first author of

```
<rdf:RDF
    xmlns:icd="http://id.who.int/icd/schema/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <rdf:Description rdf:about="http://id.who.int/icd/entity/718946808">
    <skos:closeMatch>
      <rdf:Description rdf:about="http://purl.bioontology.org/ontology/MSH/D000787">
        <skos:exactMatch>
          <rdf:Description rdf:about="http://dbpedia.org/resource/Angina_pectoris">
            <skos:definition xml:lang="en">Angina pectoris – commonly known as angina – is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm c
            <skos:definition xml:lang="en">Angina pectoris, commonly known as angina, is chest pain due to ischemia (a lack of blood, thus a lack of oxygen supply and waste remov
            <skos:definition xml:lang="en">Angina pectoris—commonly known as angina—is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm of th
            <skos:definition xml:lang="en">Angina pectoris—commonly known as angina—is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm of th
            <skos:definition xml:lang="es">La angina de pecho, también conocida como angor o angor pectoris, es un dolor, generalmente de carácter opresivo, localizado en el área
            <skos:definition xml:lang="en">Angina pectoris – commonly known as angina – is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm c
            <skos:definition xml:lang="ru">Стенокардия – заболевание, характеризующееся болезненным ощущением или чувством дискомфорта за грудиной. Боль появляется внезапно при
            <foaf:isPrimaryTopicOf rdf:resource="http://en.wikipedia.org/wiki/Angina_pectoris"/>
            <skos:definition xml:lang="en">Angina pectoris—commonly known as angina—is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm of th
            <skos:definition xml:lang="fr">L'angine de poitrine ou angor (en latin angor pectoris = « constriction de la poitrine ») est une maladie cardiaque résultant d'un manq
            <skos:definition xml:lang="zh">心絞痛是心肌缺血引起的胸痛，一般是由冠狀動脈阻塞或痙攣所導致。冠狀動脈疾病是心血管的動脈粥狀硬化，為心絞痛的主要原因。心肌缺乏血液供應時，患者會感到胸前有壓迫感。
            <skos:definition xml:lang="en">Angina pectoris – commonly known as angina – is chest pain due to ischemia of the heart muscle, generally due to obstruction or spasm c
          </rdf:Description>
        </skos:exactMatch>
        <skos:prefLabel>Angina Pectoris</skos:prefLabel>
        <skos:notation>46836/D000787</skos:notation>
      </rdf:Description>
    </skos:closeMatch>
    <skos:prefLabel>Angina pectoris</skos:prefLabel>
    <skos:definition>Angina pectoris is the result of myocardial ischemia caused by an imbalance between myocardial blood supply and oxygen demand. Angina is a common presenting
  </rdf:Description>
</rdf:RDF>
```


Figure 4. The RDF rendering of harvested textual definitions for an example ICD-11 entity Angina Pectoris

the paper (GJ) reviewed all definition entries in Chinese (n=5) available from the 10 ICD-11 entity examples, and concluded that the quality of the definitions in Chinese are reasonably good and could be useful for supporting ICD-11 multilingual definition authoring.

5 DISCUSSION

In this study, we developed a semantic web service application that provides a dynamic way to harvest textual disease definitions of Wikipedia to support the ICD-11 textual definitions authoring and its public review. The "Dynamic" means that the service application would always retrieve the most current textual definitions stored in the DBpedia dataset. We found that MeSH IDs (i.e., dbpedia:meshId) are used to code the DBpedia entries under the class "Disease", which provide a good anchor to access the textual definitions of a DBpedia entry. As of April 14, 2013, there are 5,126 entries under the class "Disease", of which 2809 (54.8%) entries have MeSH IDs annotated (covering 2505 unique IDs). In total, 19,696 (71.5%) of 27, 540 textual definitions are available for those DBpedia disease entries with MeSH IDs. In future, we will build an approach to match those DBpedia disease entries that do not have MeSH IDs coded.

To get a MeSH term mapping to a target ICD entity, we invoked the BioPortal annotation services. We used a heuristic configuration by restricting the ontology to the MeSH only and setting up the semantic types within the semantic group Disorders. In our previous study, we used the UMLS CUIs to convert the ICD-10 codes to MeSH IDs. Considering that the ICD-11 covers many new terms other than ICD-10 terms, our approach in this prototype implementation may potentially provide a better coverage though a rigorous evaluation would be needed in the future.

In addition, we used SKOS lexical and mapping properties to represent the annotations and harvested textual definitions. The main reason is that the SKOS model provides a set of semantic web friendly signatures with well-defined semantics as we demonstrated in our previous study (19).

In summary, we developed a prototype of semantic web RESTful services that automates harvesting multilingual textual definitions of Wikipedia to support ICD-11 textual definition authoring and its public review. The LexReview service application could be extended to integrate the textual definitions from other resources and subsequently consumed by a review application to support ICD-11 revision. In the future, we plan to evaluate the quality and usefulness of the harvested multilingual definitions in collaboration with WHO ICD-11 revision community.

ACKNOWLEDGEMENTS

This work was supported in part by the SHARP Area 4: Secondary Use of EHR Data (90TR000201).

Table 2. A list of ICD-11 entity examples that have Wikipedia definition matches.

ICD-11 Entity URI (Preferred Label)	Wikipedia URI (Avalilable Language Codes)	MeSH ID
http://id.who.int/icd/entity/1719064637 (Blind Loop Syndrome)	http://en.wikipedia.org/wiki/Blind_loop_syndrome (ru, en)	D005734
http://id.who.int/icd/entity/162683166 (Acute and subacute endocarditis)	http://en.wikipedia.org/wiki/Endocarditis (ru, fr, es, en)	D004696
http://id.who.int/icd/entity/761947693 (Essential (primary) hypertension)	http://en.wikipedia.org/wiki/Hypertension (zh, ru, fr, es, en)	D006973
http://id.who.int/icd/entity/925320484 (Deep vein thrombosis)	http://en.wikipedia.org/wiki/Thrombosis (es, en)	D013927
http://id.who.int/icd/entity/884453307 (Sinoatrial block)	http://en.wikipedia.org/wiki/Sinoatrial_block (fr, en)	D012848
http://id.who.int/icd/entity/1034471684 (Atrial flutter)	http://en.wikipedia.org/wiki/Atrial_flutter (fr, es, en)	D001282
http://id.who.int/icd/entity/1208831985 (Long QT syndrome)	http://en.wikipedia.org/wiki/Long_QT_syndrome (zh, fr, es, en)	D008133
http://id.who.int/icd/entity/1250136584 (Brugada syndrome)	http://en.wikipedia.org/wiki/Brugada_syndrome (zh, fr, es, en)	D053840
http://id.who.int/icd/entity/1026224967 (Lactose intolerance)	http://en.wikipedia.org/wiki/Lactose_intolerance (zh, ru, fr, es, en)	D007787
http://id.who.int/icd/entity/587895568 (Intussusception of small intestine)	http://en.wikipedia.org/wiki/Intussusception_(medical_disorder) (zh, ru, fr, es, en)	D007443

REFERENCES

1 WHO. Revision of the International Classification of Diseases (ICD). . [cited April 14, 2013]; Available from: http://www.who.int/classifications/icd/ICDRevision/en/index.html

2 WHO. ICD-11 Beta Browser. [cited April 14, 2013]; Available from: http://apps.who.int/classifications/icd11/browse/f/en

3 Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 2007 Nov;**25**(11):1251-5.

4 Jiang G, Solbrig HR, Chute CG. Using semantic web technology to support ICD-11 textual definitions authoring. ACM International Conference Proceeding Series; 2011; 2011. p. 38-44.

5 UMLS. [cited April 14, 2013]; Available from: http://www.nlm.nih.gov/research/umls/

6 SNOMED CT. [cited April 14, 2013]; Available from: http://www.ihtsdo.org/snomed-ct/

7 DBpedia. [cited April 14, 2013]; Available from: http://dbpedia.org/About

8 Wikipedia. [cited April 14, 2013]; Available from: http://wikipedia.org/

9 ICD-11 Information Model. [cited April 14, 2013]; Available from: http://informatics.mayo.edu/icd11model

10 Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on translational bioinformatics. 2009;**2009**:56-60.

11 NCBO Annotator. [cited April 14, 2013]; Available from: http://www.bioontology.org/annotator-service

12 MeSH. [cited April 14, 2013]; Available from: http://www.nlm.nih.gov/mesh/

13 The World Wide Web Consortium (W3C). [cited November 26, 2012]; Available from: http://www.w3.org/

14 SKOS. [cited April 14, 2013]; Available from: http://www.w3.org/TR/skos-primer/

15 Jersey API. [cited April 14, 2013]; Available from: http://jersey.java.net/

16 Jena ARQ API. [cited April 14, 2013]; Available from: http://jena.apache.org/documentation/query/

17 The UMLS Semantic Groups. [cited April 14, 2013]; Available from: http://semanticnetwork.nlm.nih.gov/SemGroups/

18 WHO Multilingualism. [cited Apirl 14, 2013]; Available from: http://www.who.int/about/multilingualism/en/

19 Jiang G, Solbrig HR, Chute CG. Building Standardized Semantic Web RESTful Services to Support ICD-11 Revision. ACM International Conference Proceeding Series 2012; 2012.

Flexible terminological definitions and conceptual frames

Antonio San Martín* and Pilar León-Araúz
LexiCon Research Group. Department of Translation and Interpreting, University of Granada
Calle Buensuceso n.º 11, 18071 Granada, Spain

ABSTRACT

This paper focuses on the manual creation of context-dependent natural-language definitions in EcoLexicon, a terminological knowledge base on the Environment. Given the interdisciplinary nature of the environmental domain, many concepts in EcoLexicon show a high degree of multidimensionality. In other words, this means that concepts can be described from many different perspectives. For such concepts, a single definition that encompasses the whole environmental domain is not informative enough because not all environmental domains describe concepts in the same fashion. For that reason, we propose the creation of flexible definitions.

A flexible definition is a system of definitions for the same concept composed of a general environmental definition with a set of recontextualized definitions (definitions that describe a concept from the viewpoint of a certain subject domain). This approach is based on category definitional templates and conceptual frames that provide a consistent way of managing and representing the dimensions of contextually-variable concepts in terminological definitions.

1 INTRODUCTION

A conceptual system is considered to be multidimensional when its concepts are categorized according to different characteristics, and thus showing their different dimensions (Kageura, 1997). Conceptual representations in terminological knowledge bases tend to be monodimensional. Sometimes, this may be due to the fact that the domain to be described is very constrained and there is no need to represent several dimensions. However, the usual case is that the terminologist prefers to avoid the difficulties associated with managing several dimensions. One of the problems that arise with multidimensional conceptual systems is the writing of natural-language definitions based on feature inheritance, given that the relevance of any conceptual feature can change depending on the dimension being considered and concepts can have more than one hypernym (Bowker, 1996, p. 785).

In a terminological knowledge base for translators, which is the case of EcoLexicon, the framework for this research, multidimensional knowledge representation allows the user to acquire a better insight into a given concept. This is very useful for translators because they may need to translate terms that represent concepts viewed from very different points of view.

For the representation of multidimensionality in terminological definitions, our proposal consists of the creation of several natural-language definitions for a given concept, each one describing the concept from a different subdomain of the discipline of the Environment. As a consequence, the concept to be defined is situated in different conceptual frames, which also affects which knowledge is represented in the definitions.

2 CONCEPTUAL NETWORKS IN ECOLEXICON

EcoLexicon (http://ecolexicon.ugr.es) is a terminological knowledge base on the environment. It is concept-oriented and multilingual. So far, it has 3,533 concepts and 18,798 terms in English, Spanish, German, French, Russian, Modern Greek, and Dutch as well as linguistic and phraseological information for each term. The main target users of EcoLexicon are translators, who must undoubtedly understand what they read and write in subject fields where they are not experts but need to sound like they were. This entails that they need to acquire new specialized knowledge in a very short time. To enhance knowledge acquisition, conceptual information in EcoLexicon is stored and represented in different ways.

On the one hand, specialized knowledge is represented by means of conceptual networks codified in terms of conceptual propositions in the form of a triple (CONCEPT *relation* CONCEPT), for instance, <SAND *type-of* SEDIMENT> or <MORTAR *made-of* SAND>. The conceptual relations used in EcoLexicon include both hierarchical (hypernymic and meronymic) and non-hierarchical relations, some of which are domain-specific. Concept nature (OBJECT, PROCESS, or PROPERTY) determines the combinatorial potential of concepts by means of a closed inventory of conceptual relations (León Araúz & Faber, 2010, p. 14).

On the other hand, conceptual information is also shown in the form of natural language definitions in English and Spanish, which are based on the most prototypical conceptual propositions established by the concept to be defined. Additionally, domain-specific knowledge is also presented in the form of images and videos.

2.1 Frame-based Terminology

The theoretical and methodological framework of EcoLexicon is called Frame-Based Terminology (Faber, 2012). It is a cognitive approach to Terminology inspired in the notion of frame as "any system of concepts related in such a way that to understand any one of them you have to

* To whom correspondence should be addressed: asanmartin@ugr.es

understand the whole structure in which it fits" (Fillmore, 1982).

As well as being linked by conceptual relations, concepts are also organized in the underlying conceptual framework of the Environmental Event (Fig. 1), where the most generic categories of the domain are related to each other in an action-environment interface (Faber, 2009, p. 124).

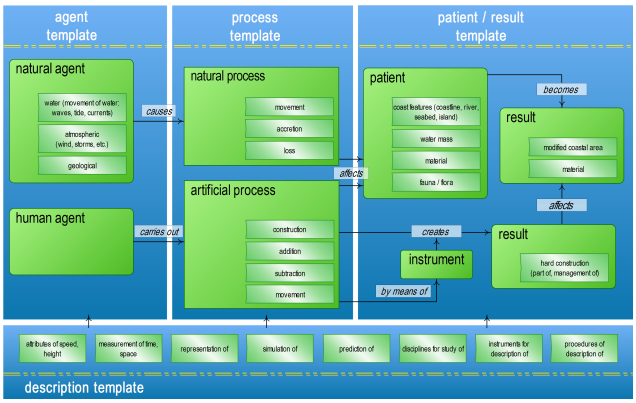

Fig. 1 Environmental Event (Reimerink & Faber, 2009, p. 631)

The configuration of the Environmental Event is mapped onto specific conceptual frames, as reflected in flexible definitions (section 4.2).

The extraction of specialized knowledge for its subsequent representation is carried out semi-automatically combining a top-down and bottom-up approach (Faber, León Araúz, & Prieto Velasco, 2009, p. 6). On the one hand, the bottom-up approach involves the extraction of information from a specialized environmental corpus compiled specifically for EcoLexicon. This is done by manual concordance analysis and keyword extraction (Tercedor & López Rodríguez, 2008), and the use of knowledge patterns (KPs) (León Araúz, Reimerink, & Faber, 2009).

On the other hand, the top-down approach consists of extracting information from the definitions of concepts in other specialized terminological resources and reference material.

2.2 Recontextualization

EcoLexicon has always focused on representing the multidimensionality of specialized knowledge, given the interdisciplinary nature of the environmental domain. However, the simultaneous representation of all conceptual dimensions often leads to an information overload in the conceptual networks (León Araúz, 2009, p. 23) (Fig. 2).

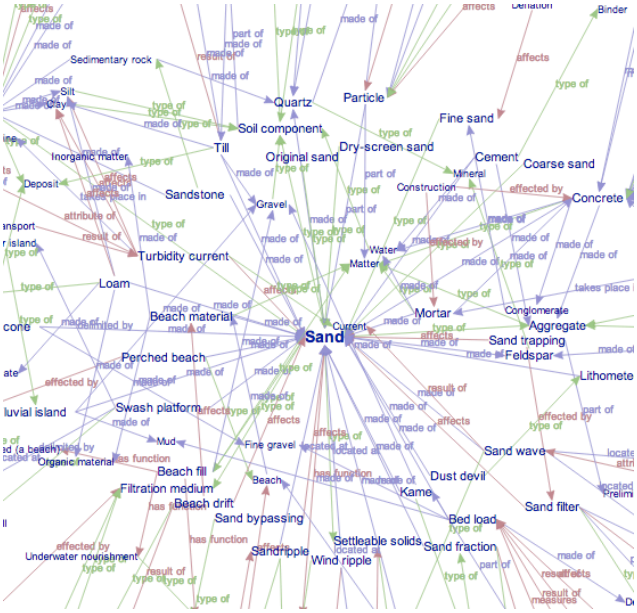

Fig. 2 Overloaded conceptual network of SAND

On Fig. 2, all the dimensions of SAND are represented at the same time. SAND is a contextually-variable concept that can have different characteristics and be linked to other concepts depending on the context. However, all the properties of the concept are never activated at the same time in real scenarios. For instance, SAND can be categorized as a subtype of SEDIMENT, a SOIL_COMPONENT, a FILTRATION MEDIUM, etc. in different contexts, but not all of them simultaneously. Contextual variation also affects other types of relation, such as functional ones. Sand, in the *CIVIL ENGINEERING* domain, is related to CONSTRUCTION, whereas in *WATER TREATMENT*, it is linked to WATER_FILTRATION.

Thus, the problem of overinformation can be resolved by recontextualizing the relational behaviour of concepts (León Araúz, 2009). Recontextualization is thus the process by which triples are stored and represented in a knowledge base, according to the restrictions imposed by different situational contexts. This approach makes it possible to account for the multidimensionality of concepts and, at the same time, to avoid information overload.

In EcoLexicon, the notion of situational context is related to the constraints imposed by environmental subdomains, such as *HYDROLOGY*, *GEOLOGY*, *OCEANOGRAPHY*, *CIVIL ENGINEERING*, etc.[1]

Recontextualization is performed by specifying which conceptual propositions (triples) are prototypically activated in each domain. This allows the representation of conceptual

[1] For the complete list of contextual domains in EcoLexicon, see León Araúz & San Martín (2011, p. 175)

networks restricted to the conceptual propositions that are salient in a certain domain (Fig. 3).

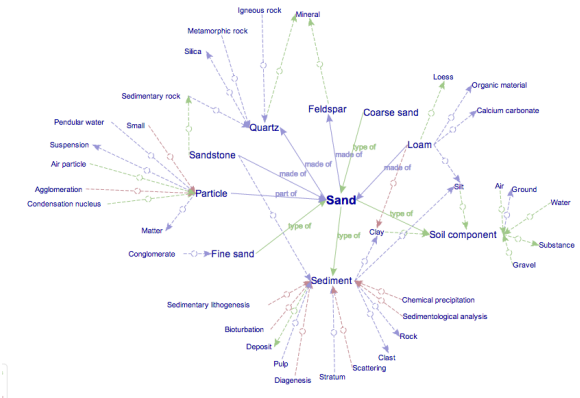

Fig. 3. SAND recontextualized in the *SOIL SCIENCES* domain.

3 DEFINITIONS IN ECOLEXICON

In EcoLexicon, definitions are regarded as mini-knowledge representations (Faber, 2002, p. 345). As such, they are based on the most representative conceptual propositions established by the concept in EcoLexicon. Each conceptual proposition is considered to be a feature of the concept.

The representativeness of each feature is determined by the category assigned to the concept being defined. Each category has a set of representative conceptual relations that describe it. They are specified in the category definitional template (León Araúz, Faber, & Montero Martínez, 2012, pp. 153–154).

3.1 Category definitional templates

Category definitional templates are schematic representations of the most prototypical relations established by the concepts that are members of the same category. They guide the formulation of definitions. They are encoded in the form of a slot-filler table like Martin's frame-based definitions (Martin, 1998). In our approach, the slots correspond to conceptual relations and the values to the concepts linked to the definiendum by means of the conceptual relations. When applying a template to a concept, it may only inherit the relation (slot) with the defined concept (value) in the template or activate a more specific concept than the one in the template. An example would be the template for HARD_COASTAL_DEFENCE_STRUCTURE (Table 1), which is applied to the definition of GROIN (Table 2), a member of this category.

HARD_COASTAL_DEFENCE_STRUCTURE	
type-of	CONSTRUCTION
located-at	SHORELINE
made-of	MATERIAL
has-function	COASTAL_DEFENCE

Table 1. HARD_COASTAL_DEFENCE_STRUCTURE category definitional template (León Araúz et al., 2012, p. 156)

GROIN	
Hard coastal defence structure made of concrete, wood, steel and/or rock perpendicular to the shoreline built to protect a shore area, retard littoral drift, reduce longshore transport and prevent beach erosion.	
type-of	HARD COASTAL DEFENCE STRUCTURE
located-at	PERPENDICULAR TO SHORELINE
made-of	CONCRETE WOOD METAL ROCK
has-function	SHORE PROTECTION LITTORAL DRIFT RETARDATION LONGSHORE TRANSPORT REDUCTION BEACH EROSION PREVENTION

Table 2. Definition of GROIN after the application of the HARD_COASTAL_DEFENCE_STRUCTURE category template (León Araúz et al., 2012, p. 156).

Category definitional templates are created by combining a bottom-up and top-down approach On the one hand, the top-down approach signifies that the membership in top-level categories partly determines the configuration of the definition. On the other hand, we also take into account the extension of a category (bottom-up approach), because a category is not only determined by its superordinates but also by its members. Consequently, before defining a concept, it is necessary to categorize it and then analyze the other members of the category so as to modulate the template inherited from superordinate categories.

4 FLEXIBLE DEFINITIONS

For concepts with a high level of contextual variation, a single definition that encompasses the whole environmental domain is not sufficiently informative, as is the case of these definitions of SAND in different environmental terminological resources:

- Mineral rock fragments (sediment) which have a particle size between 0.06 millimetres and 2.0 millimetres, which is between –1.0 and 4.0 on the phi (ϕ) scale. [A Dictionary of Environment and Conservation (Park, 2007)]
- Unconsolidated sediment consisting of mineral granules ranging between about 60 µm and 2 mm in diameter. particles of silica or quartz (SiO_2) are common components of sand. [The Environment Dictionary (Kemp, 1998)]

- A loose material consisting of small mineral particles, or rock and mineral particles, distinguishable by the naked eye; grains vary from almost spherical to angular, with a diameter range from 1/16 to 2 millimeters. [GEneral Multilingual Environmental Thesaurus (GEMET) (European Environment Agency, 2012)]

These definitions of SAND are not very useful for a translator dealing with the concept of SAND in different environmental subdomains. For instance, in *CIVIL ENGINEERING*, it is important to know the different functions of sand, and in *SOIL SCIENCES*, how sand affects the properties of the soils in which it can be found. Furthermore, no consensus seems to exist regarding SAND hypernyms (i.e. FRAGMENT, SEDIMENT, MATERIAL), because they are also source of contextual variation, which shows that knowledge is not naturally structured in clear-cut taxonomies.

For that reason, we propose the creation of 'flexible definitions'. A flexible definition is a system of definitions for the same concept composed of a general environmental definition along with a set of recontextualized definitions derived from it, which situate the concept in different domains.

Since flexible definitions follow the same premises used in the recontextualization of conceptual networks (section 2.2.), they account for the systemic factor in definition building. According to Seppälä (2012, p. 153), as a function of this factor, the relevant features to be included in a definition are determined by the conceptual system in which the concept is inserted.

Recontextualized definitions are standalone, and thus convey all the necessary information to define a concept in a certain domain, independently of the other definitions in the set. Table 3 presents an extract of the flexible definition of SAND[2]:

General Environmental Definition	Mineral material consisting mainly of particles of quartz ranging in size of 0.05-2 mm.
Geology Definition	Sediment consisting mainly of particles of quartz ranging in size of 0.05-2 mm that is part of the soil and can be found in great quantities in beaches, river beds, the seabed, and deserts.
Soil Sciences Definition	Unconsolidated inorganic soil component consisting mainly of particles of quartz ranging in size of 0.05-2 mm that are the result of weathering and erosion. It renders soils light, acidic, and permeable.
Civil Engineering Definition	Natural construction aggregate consisting mainly of particles of quartz ranging in size of 0.05-2 mm that is mixed with cement, lime and other materials to produce concrete and mortar.

Table 3. Extract of the flexible definition of SAND

In a flexible definition, the general environmental definition encodes the basic meaning present in all contextual domains and the recontextualized definitions can be considered a variation of it. For this reason, the general environmental definition includes those propositions shared by all the recontextualized definitions (e.g., in the definition of SAND: <SAND *made_of* QUARTZ>)[3].

4.1 Creation of the recontextualized hierarchies

One of the main difficulties posed by flexible definitions is that, contrary to what one might think, even hypernyms are subject to contextual variation. Quite understandably, this can impair feature inheritance in a hierarchy. As shown in Table 3, SAND is categorized in different ways depending on how the concept is prototypically conceived in each domain.

Since a coherent hierarchy needs to be specified before the defining process in order to assure correct feature inheritance[4], in the case of flexible definitions, each contextual domain requires its own hierarchy. The main information sources that determine how to categorize a concept are the definition of the concept in other terminological resources and KP-based corpus analysis.

On the one hand, extracting the hypernyms of a concept from other resources has its limitations. The first is the fact that it is not usual to find various definitions for the same concept in resources that focus on different domains. For instance, definitions of SAND can be found in Geology and Soil Sciences dictionaries and glossaries. But it is unusual to find an entry for SAND in Water Treatment or Meteorology resources, since the concept is less prototypical in the latter.

On the other hand, however, KP-based corpus analysis (Meyer, 2001) is more useful for the extraction of context-specific hypernymic relations This method permits the specification of the possible categorizations of a concept in a given contextual domain by applying KP searches to domain-specific corpora.

[2] Not all domains are included in this example.

[3] For details on how the general environmental definition is built and the way the recontextualized definitions stem from it, see León Araúz & San Martín (2012).

[4] Currently EcoLexicon is stored in the form of a relational database. Although it is in the process of becoming a formal ontology, no feature inheritance mechanisms have been implemented yet. However, terminologists manually take feature inheritance into account during conceptual modeling and definition writing.

However, all the hypernym candidates extracted with these two methods can only be used as a guide. Concepts can be categorized in several ways even in the same knowledge domain. In fact, many of the categories that can be extracted with these two methods could be considered ad hoc categories (Barsalou, 1983), constructed for a specific purpose in a certain situation and lacking conventionalization, rather than well-established categories.

The main guidelines for the structuring of recontextualized hierarchies in EcoLexicon are coherence (for correct feature inheritance once all the data is implemented in an ontology) and the activation of the most prototypical underlying conceptual frame.

4.2 Underlying conceptual frames

According to the principle of cognitive economy (Rosch, 1978, p. 28), categorization serves to mentally store and retrieve the properties generally associated with a concept in a cost-efficient manner. This also applies to the choice of genus in a definition. It follows that by choosing a genus, certain features are assigned to the definiendum (those inherited via the genus) without the need to list them explicitly in the definition.

As for recontextualized definitions, the choice of genus is even more important because by categorizing a concept as a member of a contextual domain, it is inserted into a specific conceptual frame. Such a frame takes the form of a description that relates different conceptual categories. Whereas in FrameNet (Ruppenhofer, Ellsworth, Petruck, Johnson, & Scheffczyk, 2006), frames are described by stating the relation between frame elements, in our proposal we use the categories in the Environmental Event as well as any concept in EcoLexicon. If the frame is an event composed of different stages, the information is expressed sequentially (Table 4).

Unlike in Fillmore's double-decker definitions (2003) or Maks' contextual definitions (2006), the conceptual frame is not part of the definition. It is created in order to guide the creation of the definitional templates of the categories appearing in it. In other words, the definition includes the information of the conceptual frame. As a consequence, the recontextualized definition of a concept is determined by the category to which it belongs and the underlying frame in which it takes part.

When SAND is categorized as SEDIMENT in *GEOLOGY*, SOIL_COMPONENT in *SOIL SCIENCES*, and CONSTRUCTION_AGGREGATE in *CIVIL ENGINEERING* this situates it in the frames of SEDIMENTATION (Table 4), SOIL_PROPERTIES (Table 5), and COMPOSITE_MATERIAL_PRODUCTION (Table 6), respectively.

Frame: SEDIMENTATION
Contextual domain: *GEOLOGY*
1. A MATERIAL suffers WEATHERING and EROSION and, as a consequence, becomes a SEDIMENT. 2. NATURAL_AGENTS transport the SEDIMENT. 3. The SEDIMENT is deposited in a GEOGRAPHIC_FEATURE.

Table 4. SEDIMENTATION frame

Frame: SOIL_PROPERTIES
Contextual domain: *SOIL SCIENCES*
SOIL is composed of SOIL_COMPONENTS that determine the SOIL'S PHYSICAL, CHEMICAL, and BIOLOGICAL_PROPERTIES.

Table 5. SOIL_PROPERTIES frame

Frame: COMPOSITE_MATERIAL_PRODUCTION
Contextual domain: *CIVIL ENGINEERING*
A HUMAN_AGENT produces COMPOSITE_MATERIAL by mixing a CONSTRUCTION_AGGREGATE with a MATRIX so as to use it in CONSTRUCTION.

Table 6. COMPOSITE_MATERIAL_PRODUCTION frame

5 THE CASE OF SAND IN THE CONTEXTUAL DOMAIN OF CIVIL ENGINEERING

In the *CIVIL ENGINEERING* hierarchy, SAND is categorized as a CONSTRUCTION_AGGREGATE, which itself is a subordinate of MATERIAL. Therefore, the category definitional template for MATERIAL affects the category definitional template of CONSTRUCTION_AGGREGATE, and the latter can be used, in turn, for the definition of SAND.

Table 7 and 8 show the category definitional templates for MATERIAL and CONSTRUCTION_AGGREGATE.

MATERIAL	
type-of	PHYSICAL OBJECT
made-of	SUBSTANCE
component-of	PHYSICAL OBJECT

Table.7. MATERIAL category definitional template

CONSTRUCTION_AGGREGATE	
Material consisting of particles that is mixed with a matrix to produce composite material to be used for construction.	
type-of	MATERIAL
made-of	SUBSTANCE_PARTICLES
component-of	COMPOSITE_MATERIAL (when mixed with MATRIX)

has-attribute	NATURAL/ARTIFICIAL
has-function	CONSTRUCTION

Table 8. CONSTRUCTION_AGGREGATE category definitional template with definition

The category definitional template for CONSTRUCTION_AGGREGATE is partly determined by its superordinate concept MATERIAL. Therefore, it activates the relations *made-of* and *component-of*. The underlying COMPOSITE_MATERIAL_PRODUCTION frame (Table 6) is the reason why the concepts COMPOSITE_MATERIAL and MATRIX are part of one of the values in the template. The relations *has-attribute* and *has-function*, as well as the specification that a <CONSTRUCTION_AGGREGATE is *made-of* PARTICLES>, are included in the template because of the subordinate concepts of CONSTRUCTION_AGGREGATE. An analysis of its category members such as SAND, GRAVEL, or SLAG reveals that such information is relevant for the description of the category.

Table 9 shows the definition of SAND after the application of the category definitional template of CONSTRUCTION_AGGREGATE.

SAND	
Natural construction aggregate consisting mainly of particles of quartz ranging in size of 0.05-2 mm that is mixed with cement, lime and other materials to produce concrete and mortar.	
type-of	CONSTRUCTION_AGGREGATE
made-of	(0.05-2 MM) QUARTZ_PARTICLES
component-of	MORTAR/CONCRETE (when mixed with CEMENT/LIME)
has-attribute	NATURAL

Table 9. SAND category definitional template and definition in the *CIVIL ENGINEERING* contextual domain

As can be observed in Table 9, the proposition <SAND *has-function* CONSTRUCTION> does not appear either in the definition or in the template for SAND. This is because this proposition is inherited from CONSTRUCTION_AGGREGATE, and it is thus implicit. The other relations are represented in the definition because they are a further specification of the category definitional template of CONSTRUCTION_AGGREGATE.

6 CONCLUSIONS

A single definition is not sufficient to describe multidimensional concepts that participate in many different conceptual frames. Context is an essential factor in the choice of definitional information. Depending on the context, a concept may be categorized differently and therefore establish a link to a different conceptual frame. This underlying conceptual frame guides the configuration of the category definitional template to be used in the defining process.

Because of the inherent limitations of using a closed inventory of conceptual relations, category definitional templates are not as expressive as natural language. Thus, there is the need to nuance the information in the templates. Although the configuration of category definitional templates and frames can be time-consuming, we plan to streamline these tasks in the future by formalizing all this information in an ontology. Our approach based on category definitional templates and frames provides a consistent way of managing and representing the different dimensions of contextually-variable concepts in terminological definitions. This enhances knowledge acquisition in terminological knowledge bases because it affords users a clearer and more coherent vision of each concept and its contextualized meaning in different knowledge domains.

ACKNOWLEDGEMENTS

This research was funded by the Spanish Ministry of Economy and Competitiveness (Project FFI 2011-22397) and the Spanish Ministry of Education, Culture, and Sports (FPU Program AP2009-4519).

REFERENCES

Barsalou, L. W. (1983). Ad hoc categories. *Memory cognition*, **11**(3), 211–227.

Bowker, L. (1996). Learning from Cognitive Science : Developing a New Approach to Classification in Terminology. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, and C. R. Papmehl (Eds.), *Euralex '96 Proceedings* (pp. 781–787). Göteborg: EURALEX.

European Environment Agency. (2012). GEneral Multilingual Environmental Thesaurus (GEMET). <http://www.eionet.europa.eu/gemet>

Faber, P. (2002). Terminographic definition and concept representation. In B. Maia, J. Haller, and M. Ulyrich (Eds.), *Training the Language Services Provider for the New Millennium* (pp. 343–354). Porto: Universidade do Porto.

Faber, P. (2009). The cognitive shift in terminology and specialized translation. *MonTI. Monografías de Traducción e Interpretación*, **1**, 107–134.

Faber, P. (Ed.). (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: De Gruyter Mouton.

Faber, P., León Araúz, P., and Prieto Velasco, J. A. (2009). Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies*, **1**, 1–23.

Fillmore, C. J. (1982). Frame Semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the Morning Calm* (pp. 111–137). Seoul: Hanshin.

Fillmore, C. J. (2003). Double-Decker Definitions: The Role of Frames in Meaning Explanations. *Sign Language Studies*, **3**(3), 263–295. doi:10.1353/sls.2003.0008

Kageura, K. (1997). Multifaceted/Multidimensional Concept Systems. In S. E. Wright and G. Budin (Eds.), *Handbook of Terminology Management. Volume 1: Basic Aspects of Terminology Management* (pp. 119–132). Amsterdam/Philadelphia: John Benjamins.

Kemp, D. (1998). *The Environment Dictionary*. London, New York: Routledge.

León Araúz, P. (2009). *Representación multidimensional del conocimiento especializado: el uso de marcos desde la macroestructura hasta la microestructura*. PhD Thesis. Universidad de Granada.

León Araúz, P., and Faber, P. (2010). Natural and contextual constraints for domain-specific relations. In V. Barbu Mititelu, V. Pekar, and E. Barbu (Eds.), *Proceedings of the Workshop Semantic Relations. Theory and Applications* (pp. 12–17). Valletta.

León Araúz, P., Faber, P., and Montero Martínez, S. (2012). Specialized language semantics. In P. Faber (Ed.), *A Cognitive Linguistics View of Terminology and Specialized Language* (pp. 95–175). Berlin, Boston: De Gruyter Mouton.

León Araúz, P., Reimerink, A., and Faber, P. (2009). Knowledge Extraction on Multidimensional Concepts: Corpus Pattern Analysis (CPA) and Concordances. *8ème conférence internationale Terminologie et Intelligence Artificielle*. Toulouse.

León Araúz, P., and San Martín, A. (2011). Distinguishing Polysemy from Contextual Variation in Terminological Definitions. In M. L. Carrió, J. Contreras, F. Olmo, H. Skorczynska, I. Tamarit, and D. Westall (Eds.), *Actas del X Congreso de la Asociación Europea de Lenguas para Fines Específicos: La investigación y la enseñanza aplicadas a las lenguas de especialidad y a la tecnología* (pp. 173–186). Valencia: Universitat Politècnica de València.

León Araúz, P., and San Martín, A. (2012). Multidimensional Categorization in Terminological Definitions. In R. V. Fjeld and J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 578–584). Oslo: EURALEX.

Maks, I. (2006). Frame-based definitions in a Learners' Dictionary for Dutch Business Language. In P. Ten Hacken (Ed.), *Terminology, Computing and Translation* (pp. 191–206). Tübingen: Narr.

Martin, W. (1998). Frames as definition models for terms. *Proceedings of the International Conference on Professional Communication and Knowledge Transfer. Vol .2*. (pp. 189–220). Vienna: TermNet.

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault, C. Jacquemin, and M.-C. L'Homme (Eds.), *Recent advances in computational terminology* (pp. 279–302). Amsterdam/Philadelphia: John Benjamins.

Park, C. (2007). *A Dictionary of Environment and Conservation*. Oxford, New York: Oxford University Press.

Reimerink, A., and Faber, P. (2009). EcoLexicon: A Frame-Based Knowledge Base for the Environment. In J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pillmann, I. Holoubek, and T. Bandholtz (Eds.), *European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe*. Brno: Masaryk University.

Rosch, E. (1978). Principles of Categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27–48). Lawrence Erlbaum Associates.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2006). *Framenet II: Extended theory and practice. Unpublished manuscript*.

Seppälä, S. (2012). *Contraintes sur la sélection des informations dans les définitions terminographiques: vers des modèles relationnels génériques pertinents*. PhD Thesis. Université de Genève.

Tercedor, M., and López Rodríguez, C. I. (2008). Integrating corpus data in dynamic knowledge bases The Puertoterm project. *Terminology*, **14**(2), 159–182. doi:10.1075/term.14.2.03ter

Automatic Construction of the Knowledge Base of an Onomasiological Dictionary

Gerardo Sierra*, Laura Hernández
Language Engineering Group, Engineering Institute
Universidad Nacional Autónoma de México, Ciudad Universitaria, México

ABSTRACT

For almost 14 years in the Language Engineering Group we have worked on a wide variety of Natural Language Processing (NLP) problems, being one of the earliest in the creation and operation of onomasiological dictionaries. During that time we have focused on search engine dictionary improvement, but recently our aim has been a development methodology for creating specialized onomasiological dictionaries in a semi-automatic way.

To automate the creation of onomasiological dictionaries necessarily implies the automatic execution of used processes to populate the dictionaries knowledge base. Due to the nature of these dictionaries, the definitions that must be included in the knowledge base are both normative and colloquial.

In this paper we present a proposal for semi-automatically populating the knowledge base of these dictionaries.

1 INTRODUCTION

An onomasiological dictionary is a dictionary that works in back to front way from "regular" or semasiological dictionaries. In onomasiological dictionaries users already know the definition of a term, but they do not know or have forgotten the name for that concept (this last problem is commonly known as *having a word on the tip of the tongue*) (Zock *et al*, 2011).

Onomasiological dictionaries have been classified into visual dictionaries, reverse dictionaries, thesaurus and synonym dictionaries. These dictionaries were created in order to solve the tip-of-the-tongue problem, but people still have difficulty using them because they require either that the user knows the precise words to describe the term, or its classification (i.e. when using a reverse dictionary to find the word *potato*, you might have to know that a potato is a *tuber*, and that tubers are a kind of *plant*). With visual dictionaries there is also the problem that not every concept has a visual image to represent it. For these reasons it has been suggested that free-text searcher —also known as Natural Language searching— is a viable option for solving this problem (Lancaster, 1972) since they allow the user to describe their idea of the concept in the way they would use to explain it to another human.

The creation of onomasiological dictionaries that solve inputs written in natural language improves the user experience, but it creates some major challenges that the developers must handle (Dutoit *et al,* 2002 and Bilac *et al*, 2004). The most demanding task might be the one arisen from the different ways in which a person can express the same concept, and also the fact that user definitions might not match the formal definitions found in conventional dictionaries.

In short, natural language onomasiological dictionaries need a rich knowledge base which includes not only formal, but also informal definitions. Knowledge bases can be obtained from ontologies, like in the projects Genoma KB (Cabré *et al*, 2004) and ONTODIC (Alcina, 2009). However, given the main goal of onomasiological dictionaries, for this work we decided to extract their Knowledge Bases from definitions written in texts. These definitions, on the other hand, can be used not only to populate the Knowledge Base, but also to create ontologies (Sierra, 2008).

2 DEBO

DEBO is the first onomasiological dictionary developed in the Language Engineering Group and it works with user queries given in natural language. DEBO is a specialized dictionary and it was originally made as a dictionary of Natural Disasters, but today its structure and search engine has been extrapolated to other areas such as Linguistics, Metrology, Veterinary, and Sexuality.

2.1 The search method

The dictionary works with a search engine developed by Sierra (1999) and improved later by Hernández (2011). This engine is comprised by

- A number of *terms* of an area of specialization, which are the ones that the dictionaries can retrieve as a possible answer to the user's queries.
- A *knowledge base* that includes a variety of both normative and colloquial definitions.
- A set of *key words* extracted from the definitions and associated with the terms.
- A *stop list* that contains a catalog of "empty words", such as prepositions, articles and conjunctions.

* GSierraM@iingen.unam.mx

- A set of groups of words called *paradigms*, which are groups of words with similar meaning either in area of specialization or in regular speech.

The search method follows 5 steps:

(1) The system receives the query of the user as an *input*.

(2) The system analyzes de input and extracts its keywords by filtrering them with the aid of the stop list.

(3) The system searches among the paradigms the ones to which each keyword of the input corresponds.

(4) The system searches for terms that coincide in at least one paradigm with the input's one.

(5) The system retrieves the terms ordered by the number of paradigms that each term has in common with the input —in case of a tie, the system ranks each term according to the order in which the paradigms are presented in the definition against the input—. The terms are divided in "*very probable*", "*probable*" and "*not too probable*" columns.

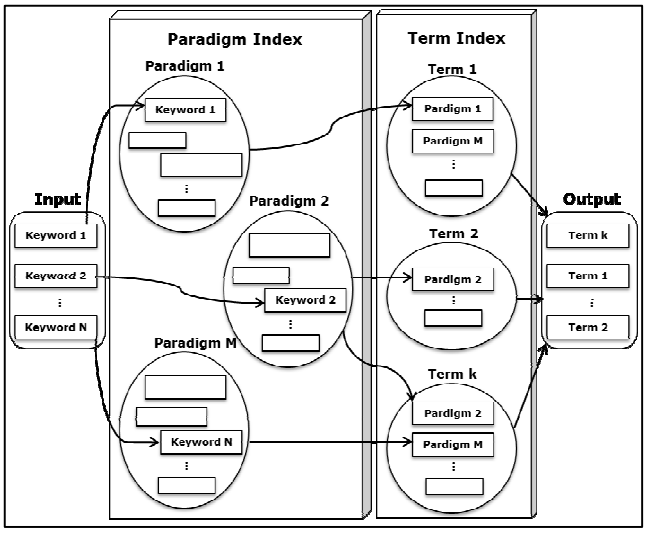

Fig. 1. Diagram of the search method.

For example, suppose someone enters as an input of the dictionary "*someone who hates gays*", and in the knowledge base there is the definition "homophobic: a person who despises homosexuals", and in the knowledge base there are also the following three paradigms.

Paradigm 1	Paradigm 2	Paradigm 3
someone	hate	gay
person	loathe	homosexual
people	contemn	lesbian
Individual	despise	queer
dude	abhor	dyke

Fig. 2. Example of paradigms in the knowledge base of an onomasiological dictionary of sexuality.

The user's definition is related to paradigms 1, 2 and 3, while one of the definitions of the term *homophobic* is related to the same paradigms in exactly the same order, which means that the term *homophobic* will be on the top of the output for this query.

2.2 The search engine performance

Hernández (2011) created the Onomasiological Dictionary of Sexuality for Mexican Spanish (DOS-MX) which used this search method. The knowledge base of this dictionary consisted of 975 both colloquial and normative definitions for 332 terms. All the definitions were found and retrieved manually from the Internet.

This dictionary had an added difficulty since it also had to be able to handle double-meaning words and phrases that are very commonly used in Mexico when talking about sex. In order to cope with this additional component, the dictionary's paradigms were extended to include double meaning words and even pejorative terms (see Paradigm 3 on Fig. 2), taking into account not only formal synonyms but also colloquial equivalents. In total there were over 33,000 different words organized in over 25,000 paradigms.

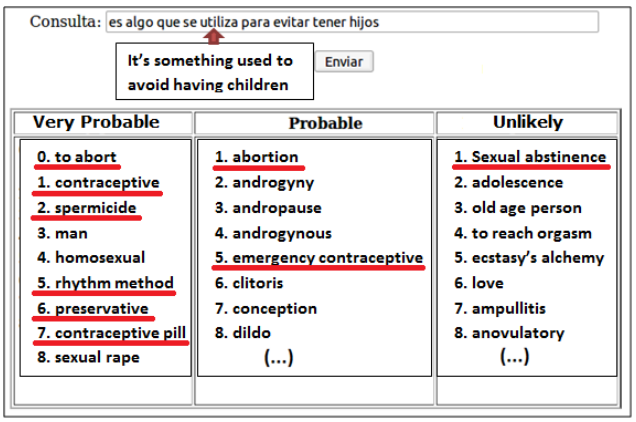

Fig. 3. Example of an output of the Onomasiological Dictionary of Sexuality for Mexican Spanish (DOS-MX)

There was an experiment to test the precision of the DOS-MX. This experiment consisted on making students write definitions of sexuality terms and to give their definitions to another student who wouldn't know which terms were described and would try to guess.

The precision of the dictionary was 71% when tested with natural language entries from users that weren't involved in the development of the dictionary, which is not bad compared to other non-English onomasiological dictionaries such as the one of El-Kahlout *et al* (2004) which has a precision of 66% in similar tests. However, a vast opportunity to improve exists.

2.3 A new improvement proposal

After the experience of the sexuality dictionary, it was concluded that the use of paradigms is not enough to try and cover all the ways in which a person can describe a concept. It was clear that there is a need to obtain many more different definitions in order to have a wide variety of expressions and ideas for every concept.

But increasing the number of definitions will also tend to increase the number of options from which the dictionary will have to choose, which is why there is also a need for organizing the definitions and terms into some sort of categories that will facilitate the selection of the correct terms.

The main problem then is to find a way to obtain a large number of definitions for the terms and classify them. This should be done in an automatized way, because by doing it manually will take too long and imply high resource usage.

3 ECODE

ECODE is a program that was developed in the Language Engineering Group with the objective of automatically detecting definitional contexts from specialized texts (Alarcón, *et al* 2008).

According to Alarcón, *et al* (2007), a definitional context is a textual fragment in which the definition of a term occurs. It is structured by a term and its definition, both being connected typographically by means of syntactic or typographic patterns.

These patterns in Spanish can be punctuation marks, such as comas, colons and parenthesis; verbs like *definir* (to define) or *significar* (to mean); discourse markers similar to *en otras palabras* (in other words), *o sea* (that is); and even pragmatic patterns like *en este conexto* (in this context) or *en términos generales* (in general terms). For example:

> *Desde el punto de vista de la sexología, se puede definir una relación sexual como el acto en el que dos personas mantienen contacto físico con el objeto de dar y/o recibir placer sexual, o con fines reproductivos.*
>
> (*From a sexology point of view, a sexual intercourse can be defined as the act in which two people have physical contact with the objective of giving and/or getting sexual pleasure or with reproductive purposes*)

The following features can be obtained from this example:

Term: "relación sexual" (*sexual intercourse*).
Definition: "acto en el que dos personas mantienen contacto físico con el objeto de dar y/o recibir placer sexual, o con fines reproductivos" (act in which two people have physical contact with the objective of giving and/or getting sexual pleasure, or with reproductive purposes).
Connecting verbal pattern: "se puede definir […] como" (can be defined as).
Pragmatic pattern as context modifier: "Desde el punto de vista de la sexología" (From a sexology point of view)

In order to automatically detect the features or components of a definitional context, Alarcón *et al* (2007) propose fifteen definitional verbal patterns divided into simple and compound ones (see Table 1).

Simple verbal definitional patterns	Compound verbal definitional patterns
• concebir (to conceive)	• consistir de (to consist of)
• definir (to define)	• consistir en (to consist in)
• entender (to understand)	• constar de (to comprise)
• identificar (to identify)	• denominar también (also denominated)
• significar (to signify)	• llamar también (also called)
	• servir para (to serve for)
	• usar como (to use as)
	• usar para (to use for)
	• utilizar como (to utilise as)
	• utilizar para (to utilise for)

Table 1. Definitional verbal patterns used by ECODE

The program processes the specialized texts and searches for definitional contexts. However, not every verbal definitional pattern that is found truly corresponds to a definition. There are some other expressions that use the same patterns with objectives other than give a definition. For this reason, Alarcón et al (2006) analyzed the use of these patterns in non-definitional contexts and found some sequences of words that are often used near a definitional verbal pattern.

Those sequences were found in some specific positions. For instance, some negation words like *no* (not) or *tampoco* (either) were found in the first position before or after the definitional verbal pattern; also adverbs like *tan* (so) as well as sequences like *no más de* (not more than) were found between the definitional verb and the nexus *como* (like); finally, syntactic sequences like adjective + verb were found in the first position after the definitional verb.

Once the system has eliminated non-definitional contexts, it proceeds to identify the features that form the definitions. For this, it uses a decision tree based on regular expressions which allows the system to identify and tag the position of every feature. These regular expressions are:

Term = BORDER (Determinant) + Noun + Adjective. {0,2} .* BORDER

Pragmatic Pattern = BORDER (sign) (Preposition | Adverb) .* (sign) BORDER

Definition = BORDER Determinant + Noun .* BORDER

The whole process of definitional contexts detection is shortened in the following diagram.

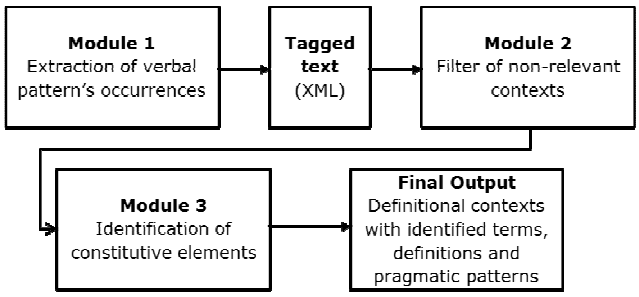

Fig. 4. ECODE architecture (taken from Alarcón, 2006)

4 DESCRIBE

ECODE was originally developed as a definitions extractor from specialized texts. However, the same definitional verbal patterns that are used in formal documents are also used in informal ones.

With this in mind, the Language Engineering Group has been working on the development of DESCRIBE, an extended scope of ECODE which extracts definitions from texts written in colloquial language.

This new adaptation consists on a module that automatically extracts search results from the Internet about a particular term, and then retrieves the contents of the web pages that match that search and analyses them looking for new definitions.

This tool removes all definitions that are repeated, and looks not only in formal websites, but also in open forums, personal webpages, blogs and chats, which provides a rich variety of definitions.

In the end, DESCRIBE retrieves a list of definition candidates that still have to be depurated, since some of the candidates might not really be definitions.

5 DEFINITION CLASSIFICATION

In order to give the dictionary search engine another feature to help the correct identification and ranking of output terms, it has been considered classifying the definitions to match not only the words and the order in which they appear, but also the type of definition given by the user and the ones in the knowledge base.

There are four kinds of definitions based on the Aristotelic definitional model: analytic, extensional, synonymic and functional (Sierra 2008). According to the LingualLinks Libray, the first one refers to "a description of the range of reference of a lexical unit" that allows readers to distinguish the term from similar words; the second kind refers to those definitions that list the objects that fall under the definition or its parts; the third kind uses synonyms or generic terms to describe the term; and finally, the fourth kind of definitions describes the term by providing its uses.

In order to automatically provide a category for each definition obtained through DESCRIBE, the verbal patterns have been divided accordingly to the kind of definition in which they usually appear.

Definition type	Verbal definitional patterns
Analytic	• concebir (to conceive) • definir (to define) • entender (to understand) • identificar (to identify) • significar (to signify)
Extensional	• consistir de (to consist of) • consistir en (to consist in) • constar de (to comprise)
Synonymic	• denominar también (also denominated) • llamar también (also called)
Functional	• servir para (to serve for) • usar como (to use as) • usar para (to use for) • utilizar como (to utilise as) • utilizar para (to utilise for)

Table 2. Definition types and their definitional verbal patterns

This definition classification is the first step in the ontology creation since, for instance, analytical definitions allow us to obtain hyponym and hypernym relations, while from extensional definitions meronymy and holonymy relations can be recovered (Soler *et al*, 2008).

6 DEFINITION CANDIDATES' DEPURATION SYSTEM

As most systems in Natural Language Processing, DESCRIBE is not perfect and sometimes the definition candidates turn out to be wrong, or the definition might be misplaced in a particular category.

The Language Engineering group has developed a tool to help the manual revision of the definition candidates' validity and their categorization correctness. This tool presents a series of definition candidates to dictionary developers. Every candidate shown to the user has also the category in which DESCRIBE placed it.

The system allows the developers to easily accept or reject a candidate as a definition and it also allows them to change the category into which the definition was originally placed.

This system helps in the task of polishing the definitions that will be part of the knowledge base of the dictionary, but it also keeps a record of the definition candidates that have been rejected. This record is intended to be used as a corpus that will serve as training data for a machine learning system that will be used to improve the precision of ECODE and, in consequence, of DESCRIBE itself.

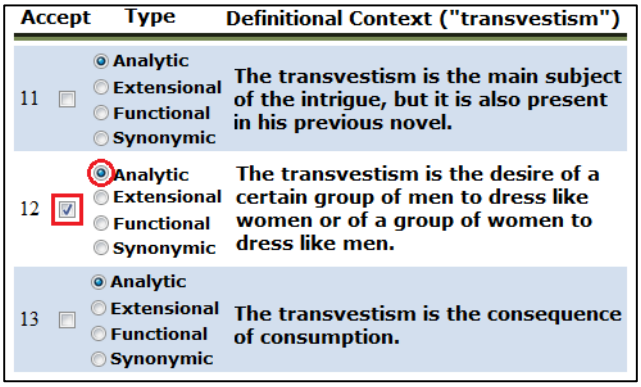

Fig. 5. Example of the use of the Definition Candidates' Depuration System.

CONCLUSIONS

The definitions included in the knowledge base of specialized onomasiological dictionaries must cover both formal and informal concepts, and they also must cover as many forms of expressing them as possible in order to procure more accurate solutions for its users.

It is also convenient to classify the definitions in the knowledge base and the ones given by the user according to their type, so as to provide the search engine with more features to compare and match the user definitions with its own, hence improving its precision. Definition classification is the first step in the creation of ontologies.

In this paper we presented a methodology to automatically obtain definition candidates to fill the knowledge base of onomasiological dictionaries and also classify these definitions according to the Aristotelic definitional model. The source of these definitions is the Internet, which allows us to a very wide variety of speakers and, for that reason, a means of expressing concepts. This methodology has been used and tested in the creation of onomasiological dictionaries of Sexuality and Linguistics, among others, and can be applied to other subject areas, such as Biomedicine, Epidemiology, Veterinary, Laws, etc..

We also presented a tool which will make possible the creation of a corpus with both good and bad definition candidates marked as such. The purpose of creating this corpus is to obtain training data for a machine learning system directed to improve the automatic detection of definitional contexts.

ACKNOWLEDGEMENTS

We would like to acknowledge DGAPA for the sponsorship of the project "Análisis estilométrico para la detección de similitud textual". We also thank the CONACYT Thematic Network "Tecnologías de la Información y la Comunicación".

REFERENCES

Alarcón, R. (2006). *Extracción automática de contextos definitorios en corpus especializados. Propuesta para el desarrollo de un ECCODE (extractor de candidatos a contextos definitorios)*. Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra, Barcelona (Doctoral thesis):

Alarcón, R., Bach, C., and Sierra, G. (2008). *Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica*. Revista Española de Lingüística **37**, 247-278.

Alarcón, R., and Sierra, G. (2006). *Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios*. Avances en la Ciencia de la Computación, VII Encuentro Nacional de Ciencias de la Computación, 242-247.

Alarcón, R., Sierra, G., and Bach, C. (2007). *Developing a Definitional Knowledge Extraction System*. Proc. 3rd Language and Technology Conference (L&TC'07), Adam Mickiewicz University, Poznan, Polonia.

Alcina, A. (2009). *Metodología y tecnologías para la elaboración de diccionarios terminológicos onomasiológicos*. Terminología y sociedad del conocimiento. Bern: Peter Lang, 33-58.

Bilac, S., Watanabe, W., Hashimoto, T., Tokunaga, T. y Tanaka, H. (2004). *Dictionary search based on the target word description*. Proceedings of the Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004), 556-559.

Cabré, M. T., Bach, C., Estopà, R., Feliu, J., Martínez, G., and Vivaldi, J. (2004). *The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities*. 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisboa, European Languages Resources Association, 87-90.

Dutoit, D. y Nugues, P. (2002). *A Lexical Database and an Algorithm to Find Words from Definitions*. Proceedings of the 15th European Conference on Artificial Intelligence, Lyon, 450-454.

El-Kahlout, I., and Oflazer, K. (2004). *Use of Wordnet for Retrieving Words from Their Meanings*. 2nd Global WordNet Conference, Brno, Czech Republic.

Hernández, L. (2011). *Creación semi-automática de la base de datos y mejora del motor de búsqueda de un diccionario onomasiológico*. Universidad Autónoma de México (Master thesis).

Lanacaster, F (1972). *Vocabulary control for information retrieval*. Washington: Information Resources Press.

Sierra, G. (1999). *Design of a concept-oriented tool for terminology*. UMIST, Manchester (Doctoral thesis).

Sierra G., Alarcón R., Aguilar C., and Bach C. (2008). *Definitional verbal patterns for semantic relation extraction*. Terminology 14(1), pp. 74-98.

Soler, V., and Alcina, A. (2008). *Patrones léxicos para la extracción de conceptos vinculados por la relación parte-todo en español*. Terminology 14(1).

Zock, M., and Rapp Reinhard (2011). *Introduction to this special issue on Cognitive Aspects of Natural Language Processing*. Journal of Cognitive Science 12(3).

Introduction to the Logic of Definitions

Barry Smith

What follows is a summary of basic principles pertaining to the definitions used in constructing an ontology. A definition is a statement of necessary and sufficient conditions. What this means in the simplest case can be understood as follows. To say that ϕ-ing is a *necessary condition* for being an A is just another way of saying that every A ϕ's; to say that ϕ-ing is a *sufficient condition* for being an A is just another way of saying that everything that ϕ's is an A. The goal in writing a definition is to specify a set of conditions of this sort which are all necessary, and which are jointly sufficient.

The following is a set of necessary conditions for being a triangle which are also jointly sufficient, and which thus form a definition:

> *X* is a triangle =def. *X* is a closed figure; *X* has exactly three sides; each of *X*'s sides is straight; *X* lies in a plane.

Everything which satisfies all of the conditions on the right hand side is also a triangle. And everything which is a triangle satisfies all of these conditions.

Not every statement of necessary and jointly sufficient conditions is a definition. 1. The statement of necessary and sufficient conditions used to define the term A should itself use terms which are easier to understand than (and are logically simpler than) the term A itself. 2. The necessary and sufficient conditions must be satisfiable; that is, there must be actual examples of entities which satisfy the definition. Thus we cannot, for example, define a perpetual motion machine as a prime number that is divisible by 4, even though everything which is a perpetual motion machine is also a prime number that is divisible by 4.

A useful template for creating definitions along the lines described above is provided by what are called Aristotelian definitions, which is to say definitions of the form

> S =def. a G which Ds

where 'G' (for: genus) is the parent term of 'S' (for: species) in some ontology. Here 'D' stands for 'differentia', which is to say that 'D' tells us what it is about certain Gs in virtue of which they are Ss. An example Aristotelian definition (from the Foundational Model of Anatomy Ontology):

> cell =def. an anatomical structure which consists of cytoplasm surrounded by a plasma membrane
>
> plasma membrane =def. a cell part that surrounds the cytoplasm

The benefits of using Aristotelian definitions are 1. That each definition reflects the position in the ontology hierarchy to which the defined term belongs. Every definition, when unpacked, takes us back to the root node of the ontology to which it belongs. 2. Circularity is prevented automatically. 3. The definition author always knows where to start when formulating a definition. 4. It is easier to coordinate the work of multiple definition authors.

Aristotelian definitions work well for common nouns (and thus for the names of types or universals by which ontologies are principally populated). They do not work at all for those common nouns which are in the root position in an ontology, for here there is no parent term (no genus) to serve as starting point for definition. Root nodes in an ontology must therefore either be defined using as genus some more general term taken from a higher-level ontology such as BFO, or they must be declared as primitive. Primitive terms cannot be defined, but they can be elucidated (by means of illustrative examples, statements of recommended usage, and axioms).

Note that the Aristotelian rule will bring the benefits mentioned above only if the ontology in question satisfies the principle of single inheritance, which is to say, only if each term in the ontology has at most one parent. For only thus is the choice of 'G' for each given 'S' unique. Single inheritance itself however brings multiple benefits to ontology authoring: 1. It prevents a number of common errors which derive from the tendency once dominant among ontology authors of what has been called "is-a overloading". 2. It promotes integration of an ontology with its neighboring ontologies. 3. It promotes forking of ontologies. 4. The benefits of multiple inheritance, for example in terms of surveyability of an ontology (so that it is easier for human beings to find the terms they need) can be gained in any case by formulating the official (or 'asserted') version of an ontology as an asserted monohierarchy and allowing the development of inferred polyhierarchies for specific groups of users.

References

Cornelius Rosse, J. Leonardo V. Mejino Jr., The Foundational Model of Anatomy Ontology. In: Burger, A. Davidson, Baldock R. (eds), *Anatomy Ontologies for Bioinformatics: Principles and Practice*, (2007) 59-117, London: Springer.

Barry Smith and Werner Ceusters, "Ontological Realism as a Methodology for Coordinated Evolution of Scientific Ontologies", *Applied Ontology*, 5 (2010), 139–188. PMC3104413

Semantic Trilogy '13

Vaccine and Drug Ontology Studies

Montreal, Quebec, Canada July 6th—12th 2013

UNIVERSITÉ Concordia UNIVERSITY

International Workshop in Vaccine and Drug Ontology Studies (VDOS 2013)

Proceedings

edited by

Cui Tao,
Yongqun (Oliver) He,
Luca Toldo, and
Sivaram Arabandi

VDOS 2013 is held in conjunction with the
4th International Conference on Biomedical Ontology
(ICBO 2013)

Concordia University
Montreal, Quebec, Canada

July 7, 2013

Preface

The VDOS international workshop series focuses on vaccine- and drug-related ontology modelling and applications. Drugs and vaccines have contributed to dramatic improvements in public health worldwide. Over the last decade, tremendous efforts have been made in the biomedical ontology community to ontologically represent various areas associated with vaccines and drugs – extending existing clinical terminology systems such as SNOMED, RxNorm, NDF-RT, and MedDRA, as well as developing new models such as Vaccine Ontology. The VDOS workshop series provides a platform for discussing innovative solutions as well as the challenges in the development and application of biomedical ontologies for representing and analysing drugs and vaccines, their administration, host immune responses, adverse events, and other related topics.

The VDOS-2013 workshop is the 2nd in this series and takes place in Montreal, QC, Canada, on July 7th, 2013, in conjunction with the 4th International Conference on Biomedical Ontology (ICBO 2013). The first workshop of the series was organized as the "Vaccine and Drug Ontology in the Study of Mechanism and Effect" workshop (VDOSME 2012) on July 21, 2012, at Graz, Germany, as part of the third International Conference on Biomedical Ontology (ICBO 2012). For this year, the name has been changed to *"Vaccine and Drug Ontology Studies (VDOS)"* to reflect the expansion in the scope to more than just mechanism and effect. The workshop series also covers vaccine and drug-related clinical data representation and analysis, including clinically reported vaccine and drug adverse events.

The co-organizers of VDOS-2013 include:
Cui Tao, Yongqun (Oliver) He, Luca Toldo, and Sivaram Arabandi

Standardized Drug and Pharmacological Class Network Construction

Qian Zhu[1] *, Guoqian Jiang [1], Liwei Wang[2], Christopher G. Chute[1]
[1] Department of Health Sciences Research,
Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA
[2] School of Public Health, Jilin University, Changchun, Jilin, China

ABSTRACT

Dozens of drug terminologies and resources capture the drug and/or drug class information, ranging from their coverage and adequacy of representation. No transformative ways are available to link them together in a standard way, which hinders data integration and data representation for drug-related clinical and translational studies. In this paper, we introduce our preliminary work for building a standardized drug and drug class network that integrates multiple drug terminological resources, using Anatomical Therapeutic Chemical (ATC) and National Drug File Reference Terminology (NDF-RT) as network backbone, and expanding with RxNorm and Structured Product Label (SPL). In total, the network consists of 39,728 drugs and drug classes. Meanwhile, we calculated and compared structure similarity for each drug / drug class pair from ATC and NDF-RT, and analysed constructed drug class network from chemical structure perspective.

1 INTRODUCTION

Drug classes are group names for drugs that have similar activities or are used for a same type of disease and disorder. There are different ways to classify drugs. One way is to group drugs based on their therapeutic use or class (e.g., antiarrhythmic or diuretic drugs) as used by Anatomical Therapeutic Chemical (ATC) [1]. Another way is to group drugs using their dominant mechanism of action as used by National Drug File Reference Terminology (NDF-RT) [2]. However, drug classes defined by different systems are not compatible. It is worth to compare and integrate them in a universal fashion in order to support clinical related studies better. For example, Mougin, et al. [3] conducted a study for comparing drug classes between ATC and NDF-RT focusing on the relations between drugs and pharmacological classes (i.e., drug-class membership relations), which will facilitate the integration of these two resources.

Drug terminologies define drug entities as well as relevant properties and relationships with pharmacological classes. Drug terminologies are usually developed and maintained by different institutions using site-specific drug coding systems. Heterogeneous drug representations across different systems make it difficult to navigate diverse drug resources. The lack of a transformative way to link heterogeneous drug resources hinders data integration and data representation for drug-related clinical and translational studies. To overcome this obstacle, we proposed to represent drug information from diverse resources in a standard and integrated manner.

ATC and NDF-RT are the proposed sources of drug classification information. In the present study, we developed an approach to map drug and drug class entities from ATC and NDF-RT to UMLS (Unified Medical Language System) [4] and generated these mappings as a drug network backbone. Furthermore, we extended such network with RxNorm [5] and Structured Product Labeling (SPL) [6] integration, benefited from the broad drug relevant knowledge provided by these two resources. RxNorm provides links among different vocabularies, e.g. NDF-RT. SPL contains full drug interaction information, such as drug and drug interaction, and adverse drug event, etc., which has been explored and implemented by investigators and relevant applications have been developed, such as LinkedSPLs [7], ADEPedia [8]. Additionally, to extend and compare the drug classes defined by ATC and NDF-RT from chemical structure point of view, we introduced chemical structure similarity with the assumption that similar molecules have similar activities.

The paper is organized in several sections. We introduce the background knowledge about the resources and tools used in material section; in the methods section, we introduce the workflow details for network construction; then followed by discussion and conclusion sections.

2 MATERIALS AND METHODS

NDF-RT is a well-known drug terminological resource, and snapshot of NDF-RT was downloaded as of Nov. 8, 2012. In ATC classification system, drugs are categorized into different groups at five different levels according to the organ or system on which they act and/or their therapeutic and chemical characteristics [9]. ATC with a released version on January 2012 was used in this study. RxNorm provides normalized names for clinical drugs and links them to several drug vocabularies differentiating by "SAB" label. For example, "SAB=MTHSPL" indicates the source from SPL and "SAB=NDFRT" from NDF-RT. Two files are used in this study: 1) RXNCONSO.RRF, including all connections with source vocabularies. 2) RXNREL.RRF including relationships among concepts. RxNorm used in this study was

* To whom correspondence should be addressed: zhu.qian@mayo.edu

the version of Oct. 2012. SPL contains structured content of labeling (all text, tables and figures), along with additional machine readable information. The mappings between SPL and RxNorm used in this study are extracted from RxNorm RXNCONSO files with SAB = MTHSPL.

In this paper, we introduce a drug and drug class network by utilizing multiple drug terminological resources: ATC, NDF-RT, RxNorm, and SPL. ATC and NDF-RT are used as the network backbone, from which we integrated RxNorm and SPL as extension. Meanwhile, we calculated structure similarity for drug pairs from ATC and NDF-RT, and clustered them by structural similarity. The details of each step conducted in this study are described in the following sections.

2.1 Mapping NDF-RT with ATC

To map NDF-RT with ATC via UMLS, we translated NUI, NDF-RT Numerical Unique Identifier, and ATC name to UMLS CUI, UMLS concept unique identifier.

3.1.1 ATC mapping to UMLS

ATC is not well integrated with other drug terminologies (e.g., NDF-RT), as it uses its own coding system to code drug entities. To map ATC with NDF-RT and present the drug network transformatively by using standard representation, UMLS, we employed NCBO annotator [10] to semantically annotate each ATC name. Among more than 200 ontologies from UMLS Metathesaurus and NCBO BioPortal [11], RxNorm and NDF-RT have higher priority in this study. To avoid unnecessary annotations by non-drug relevant ontologies, we limited UMLS semantic types [12] within "Chemicals & Drugs" semantic group [13]. We extracted ontology id and concept id, which are two mandatory input parameters to invoke NCBO BioPortal REST API [14] for searching UMLS CUI, from the annotation results.

3.1.2 NDF-RT mapping to RxNorm and UMLS

NDF-RT concepts are organized into different categories with corresponding category labels. For example, "N0000179008, 1,1,1-trichloroethane, [Chemical/Ingredient]" and "N0000175641, Autonomic Ganglionic Blocker, [EPC]" are chemical ingredient and EPC class respectively. In this study, we retrieved the concepts that are labeled as VA class, VA product, EPC, Chemical ingredient and generic ingredient combination.

SQL query was executed to search RxCUIs (RxNorm Concept Unique Identifier) from RxNorm RXNCONSO table that was pre-loaded into our local MySQL database for NUIs. We retrieved UMLS CUI by invoking NLM RxNav RESTful API [15] with each NUI as an input parameter.

2.2 Calculating structure similarity

To analyze and expand the drug and drug class network from chemical structure perspective, we calculated the structure similarity among the drug pairs from ATC and NDF-RT, and grouped them using the score of structure similarity as Tanimoto Coefficient, i.e., similarity between these pairs of descriptors [16]. The cutoff value of the structure similarity is set as the score greater than 0.85, as it exhibits similar biological activity between the two molecules.

We first converted NDF-RT drug name and ATC name to SIMILES (Simplified molecular-input line-entry system) [17] as chemical representation by invoking PubChem entrez web service [18] and NCI resolver [19] REST API. Then we translated SMILES to chemical fingerprint and calculated Tanimoto similarity by using the aforementioned CDK functions.

2.3 Integrating RxNorm and SPL mappings

Mappings among RxNorm, SPL and NDF-RT are provided by RxNorm and available in the RxNorm RXNCONSO table. Two steps were performed to retrieve these mappings. First, we obtained concepts labeled as "SAB=NDFRT" and "SAB=RXNORM", denoted as RxNorm and NDF-RT mappings. Then, we searched for the concepts with "SAB=MTHSPL" label from the concepts identified in the first step. Then the final list of concepts is the common concepts across the three resources.

The network has been expanded from NDF-RT nodes that have mappings with RxNorm and SPL. We extracted SPL identifier (setId) from RXNREL table and saved for future SPL relevant information, LinkedSPL integration.

In addition, we performed a case study to demonstrate the usefulness of the drug and drug class network.

3 RESULTS

There are total 5,717 individual entities, which correspond to 4,483 distinct ATC names, i.e. one drug can be categorized into multiple therapeutic classes (more details described in the Discussion section).

Of 48,266 NDF-RT concepts, 34,011 concepts were used in this study, consisting of 15,857 VA Products, 486 VA classes, 9,960 Chemical/Ingredients, 7,184 Generic Ingredient Combinations, and 524 EPC. The child and parent relationships among these NDF-RT concepts are retrieved and stored from RxNorm RXNREL table via "CHD" (concept 1 is a child of concept 2) and "PAR" (concept 1 is a parent of concept 2) labels.

RxNorm, SPL and NDF-RT mappings were extracted from two RxNorm files: RXNCONSO and RXNREL, which were loaded into MySQL database.

3.1 Results for ATC and NDF-RT mappings

In order to build drug and drug class network with ATC and NDF-RT as backbone, first of all, we mapped ATC entities with NDF-RT concepts via UMLS, four steps involved.

4.1.1 ATC Annotated by NCBO

3,607 ATC entities including 3,152 drugs and 455 drug classes were mapped to UMLS CUIs by two ontologies, RxNorm and NDF-RT from NCBO BioPortal. Of these 3607 ATC mappings, 2180 ATC entities were exactly matched with the preferred names from RxNorm and NDF-RT. 866 ATC entities including 211 drug classes and 655 drugs were mapped to other ontologies available from NCBO. There are 1,244 ATC entities (21.8%) including 657 drugs and 587 drug classes failed to map to UMLS due to no annotations generated accordingly. We attempted to map these failed ATC names with RxNorm directly by invoking NLM RxNav RESTful API [20] with ATC names as input parameter, but none of them got mapping results. The failure reasons are discussed in the discussion section further.

4.1.2 NCBO annotation evaluation

The annotations were automated programmatically using NCBO Annotator Web Services API. We manually evaluated the annotation results. Of the 4,473 annotations with NDF-RT and RxNorm, 2,401 exact mappings were not further evaluated. The authors (QZ, LW) manually reviewed the rest of annotations (2,072 in total). As the evaluation results, 88.7% is correct, 10.3% is partial mappings, and 1.0% is incorrect. The precision was calculated as 99.5%, recall as 78.2% and F-measure as 87.4%, in which we counted exact mappings, partial mappings and correct mappings (4,453 in total) as true positive, 1,244 failed mappings as false negative and 20 incorrect mappings as false positive.

4.1.3 Mapping NDF-RT to RxNorm and UMLS

NDF-RT and RxNorm mappings exist in the RXNCONSO table with "SAB=NDFRT" label. Consequently, RxCUI corresponding to each NDF-RT concept can be retrieved from these mappings directly.

NDF-RT provides UMLS mappings. Hence, to retrieve UMLS for each NDF-RT concept, we called NLM NDF-RT RESTful API [9]. The searching results are shown in Table 1. 99.2% NDF-RT concepts have been mapped to UMLS.

NDF-RT Concepts	NUI	UMLS CUI
Chemical/Ingredient (9,960)	9,934	9,932
VA Class (486)	486	483
VA Product (15,857)	15,695	13,263
EPC (524)	480	478
Generic ingredient combination (7,184)	7,139	6,801
Total (34,011)	33,734	30,957

Table 1. UMLS CUI retrieval by RxNav NDF-RT API

4.1.4 ATC and NDF-RT mapping

In total, 3,850 distinct mappings between ATC and NDF-RT were generated, including 2,015 chemical/ingredients, 1,826 Generic Ingredient Combinations and 1 VA class. It includes distinct 2,226 ATC entities, covering 99 drug classes, and 2,127 individual drugs.

3.2 Results for structural similarity calculation

SMILES have been retrieved for all drugs from ATC and NDF-RT via PubChem Entrez web API and NCI Resolver web API. 2,618 ATC entities have gotten SMILES from NCI, 3,471 entries retrieved from PubChem. Combining NCI and PubChem searching results, total 3,487 ATC entries got SMILES, and 9,132 unique NDF-RT concepts got SMILES.

We calculated the Tanimoto coefficient as structure similarity for each pair of concepts from ATC and NDF-RT separately by converting SMILES to fingerprint. Then we got 8,513 pairs from ATC and 69,882 pairs from NDF-RT with Tanimoto coefficient greater than 0.85, and integrated them into the drug and drug class network.

3.3 Results for NDF-RT, RxNorm and SPL mapping

We integrated RxNorm and SPL mappings with NDF-RT. The mappings between RXNORM, NDF-RT and SPL resulted in 5,838 unique RxNorm concepts with 36,408 NDF-RT concepts and 41,188 SPL labels. The mappings mostly fall into two main categories according to term types defined by RxNorm, 3,056 are Semantic Clinical Drugs and 1,543 are Ingredients.

It is worthy to note that one RxNorm concept may be mapped to multiple NDF-RT and/or SPL concepts, for example, RxCUI "74" mapped to 3 NUIs in NDF-RT including N0000006481, N0000147349, N0000006481 and 11 set_ids in MTHSPL such as 0d65128b-8eb7-440b-870a-7e3be18152b3,1e6d6cd5-ab14-4258-a0fe-5f6a3cae437f.

4 DISCUSSION

In this study, we successfully built a drug and drug class network with 39,728 concepts from ATC and NDF-RT. All concepts were mapped to UMLS and labeled as UMLS CUIs accordingly. We also integrated RxNorm and SPL mappings, and extended the network with structure similarity calculation.

4.1 ATC to UMLS mapping

In total, 77.9% ATC terms have been mapped to UMLS. Comparing to 68.7% mapping results conducted by Merabti et al [21], our study shows the improvement of mappings from ATC to UMLS by leveraging NCBO annotator. However, 22.1% ATC terms failed to be mapped due to several reasons as follows, 1) Many of the ATC terms are combinations of multiple concepts, such as "calcium acetate and magnesium carbonate", "combinations of sulfonamides and trimethoprim, including derivatives"; 2) The exclusions are embedded in the ATC names, such as "platelet aggregation inhibitors excluding heparin", "nutrients without phenylalanine; 3) Non-standard representation is used by ATC though we corrected and expanded some abbreviations occurring in

ATC name. For example, "DIGESTIVES, INCL. ENZYMES" was corrected to "DIGESTIVES, INCLUDING ENZYMES"; 4) Non-drug terms are used, especially for drug classes in ATC, such as "VARIOUS", "SENSORY ORGANS". Above obstacles were the main reasons for mapping failure. In the future study, we will explore MMTx program that reported by Mougin et al. [3], and more NLP (Nature Language Processing) algorithms to parse ATC names for improving the mapping performance between the ATC and the UMLS.

4.2 Benefits from structure similarity integration

Structure similarity calculation applied in this study enables connections among the drug nodes sharing common similar chemical substructures. Beside the benefit shown in the case study, this integration also provides relevant clues for guiding clinical decision support system from the structure perspective as it offers a full profile of therapeutics for individual drugs. ATC classification system categorizes drugs according to its therapeutic classes; hence, one ATC drug can be grouped into multiple categories due to its diverse therapeutic functionalities. For instance, "Thonzylamine" is an antihistamine and anticholinergic used as an antipruritic and is grouped into two categories: "antiallergic agents" and "antihistamines for topical use" within the ATC hierarchy. The corresponding two ATC entities (R01AC06 and D04AA01) for "Thonzylamine" in two separate classes ("R" and "D") are connected based on similarity score that is equal to 1. Thus, the entities within these two categories are connected, and physicians would be able to utilize such knowledge for Thonzylamine for their clinical decision from both therapeutics and structure point of view.

4.3 Future work

Drug entity mapping algorithm will be modified to enable more connections detected; more human review will be expected to improve the accuracy of the mappings. Meanwhile, we will seek possible collaborations with external sites such as the NLM for improving such mapping algorithm development. We will integrate more drug related resources, such as Drugbank and PharmGKB, and drug interaction data, drug and adverse event data as shown in Figure 1. The entire data set generated in this project will be released to public once the proposed action items accomplished.

5 CONCLUSION

We successfully integrated NDF-RT, ATC, RxNorm and SPL and built a drug and drug class network using standardized identifier for representing drug and drug class entities. In addition, the network was expanded from chemical structure perspective by similarity calculation. More other drug terminological resources and drug interaction information will be integrated in the future study.

ACKNOWLEDGMENTS

This work was supported by the Pharmacogenomic Research Network (NIH/NIGMS-U19 GM61388) and the SHARP Area 4: Secondary Use of EHR Data (90TR000201).

REFERENCES

[1] ATC: http://www.who.int/classifications/atcddd/en/. Accessed by Apr.11.2013.
[2] NDF-RT: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/. Accessed by Apr.11.2013.
[3] Mougin, F., Burgun, A., and Bodenreider, O. Comparing Drug-Class Membership in ATC and NDF-RT . Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012:437-443.
[4] Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004, 32, 267–270
[5] RxNorm: www.nlm.nih.gov/research/umls/rxnorm. Accessed by Apr.11.2013.
[6] Structured Product Labeling: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm. Accessed by Apr.11.2013.
[7] Hassanzadeh O, Zhu Q, Freimuth R, Boyce R, Extending the "Web of Drug Identity" with Knowledge Extracted from United States Product Labels, submitted to AMIA Summit on Clinical Research Informatics, 2013
[8] Jiang G, Solbrig H. R, Chute C.G. ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology. AMIA Annu Symp Proc. 2011:607-16.
[9]http://en.wikipedia.org/wiki/Anatomical_Therapeutic_Chemical_Classification_System. Accessed by Apr.11.2013.
[10] Jonquet C., Shah N., Musen M. The Open Biomedical Annotator. AMIA Summit on Translational Bioinformatics; 2009: 56–60. The NCBO Annotator web service: http://www.bioontology.org/annotator-service. Accessed by Apr.11.2013.
[11] Noy, N., Shah, N., Dai, B., Dorf, M., Gri_th, N., Jonquet, C., Montegut, M., Rubin, D., Youn, C., Musen, M.: Bioportal: A web repository for biomedical ontologies and data resources. In: Demo session at 7th International Semantic Web Conference (ISWC 2008)
[12] Semantic Type: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html. Accessed by Apr.11.2013
[13] Bodenreider O, McCray AT Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 2003; 36(6):414-432.
[14] BioPortal REST services: http://www.bioontology.org/wiki/index.php/NCBO_REST_services. Accessed by Apr.11.2013.
[15] NDF-RT RESTful API: http://rxnav.nlm.nih.gov/NdfrtRestAPI.html#label:r24. Accessed by Apr.11.2013.
[16] Holliday JD, Hu CY, Willett P, Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bitstrings. Comb Chem High Throughput Screen, 2002, 5(2):155-66.
[17] SMILES: http://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system. Accessed by Apr.11.2013.
[18] PubChem Entrez: http://www.ncbi.nlm.nih.gov/books/NBK25500/. Accessed by Apr.11.2013.
[19] NCI resolver: http://cactus.nci.nih.gov/chemical/structure. Accessed by Apr.11.2013.
[20] RxNorm RESTful API: http://rxnav.nlm.nih.gov/RxNormRestAPI.html. Accessed by Apr.11.2013.
[21] Merabti et al, 2011, Stud Health Technol Inform. 2011;166:206-13

Panacea, a Semantic-enabled Drug Recommendations Discovery Framework

Charalampos Doulaverakis [1]*, George Nikolaidis [2], Athanasios Kleontas MD [2,3] and Ioannis Kompatsiaris [1]

[1]Centre for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece
[2]Ergobyte S.A., Thessaloniki, Greece
[3]Theagenio Cancer Hospital, Thessaloniki, Greece

ABSTRACT

The paper presents Panacea, a semantic framework capable of offering drug-drug and drug-diseases interaction discovery. For enabling this kind of service, medical information and terminology had to be translated to ontological terms and be appropriately coupled with medical knowledge of the field. International standards, such as the ICD-10 and ATC classifications, provide the backbone of the common representation of medical data while the medical knowledge of drug interactions is represented by a rule base which makes use of the aforementioned standards. Representation is based on the lightweight SKOS ontology. A layered reasoning approach is implemented where at the first layer ontological inference is used in order to discover underlying knowledge, while at the second layer a two-step rule selection strategy is followed resulting in a computationally efficient reasoning approach. Details of the system architecture are presented while also giving an outline of the difficulties that had to be overcome. The paper compares the current approach to a previous published work by the authors, a service for drug recommendations named GalenOWL, and presents their differences in modelling and approach to the problem, while also pinpointing the advantages of Panacea.

1 INTRODUCTION

One of the health sectors where intelligent information management and information sharing compose valuable preconditions for the delivery of top quality services is personalized drug prescription. This is more evident in cases where more than one drug is required to be prescribed, a situation which is not uncommon, as drug interactions may appear. The problem is magnified by the wide range of available drug substances in combination with the various excipients in which the former are present.

If one takes into account that there exist more than 18,000 pharmaceutical substances, including their excipients, then it is clear that the continuous update of health care professionals is remarkably hard. Over this, the extensive literature makes discovery of relevant information a time consuming and difficult process, while the different terminologies that appear between sources add more burden on the efforts of medical professionals to study available information.

Semantic Web technologies can play an important role in the structural organization of the available medical information in a manner which will enable efficient discovery and access. Research projects funded for enabling Semantic Web technologies in the diagnosis and therapeutic procedures exist such as PSIP (Beuscart *et al*., 2009) and Active Semantic Documents (Sheth, 2005) or works such as (Adnan *et al*., 2010), but they don't fully address the problem of automated drug prescription using drug-drug and drug-disease interactions.

Rule-based approaches have been proposed for addressing issues relating to biomedical ontologies research. It is common for ontologies written in expressive Semantic Web languages such as OWL, not be able to handle all requirements for capturing the knowledge in several biomedical and medicine domains. As a method for enriching the expressiveness of ontology languages, researchers have proposed the use of rules which act upon the defined ontological knowledge. According to (Golbreich, 2004) rules are helpful in the following situations relating to biomedical ontologies: defining "standard rules" for chaining ontology properties, "bridging rules" for reasoning across different domains, "mapping rules" for defining mappings between ontologies entities and "querying rules" for expressing complex queries upon ontologies. The author gives a thorough review of RuleML and SWRL, the two major ontology rule languages, the available rule formation tools and the reasoners. (Golbreich *et al*., 2005) make use of the outcomes of the previous paper to showcase the need for rules in biomedical applications with a use case of a brain anatomy definition, where a brain structure ontology is defined in OWL but rules describing the relationships between the properties and entities are needed for correct annotation of MRI images. Another work citing the need for semantically enriched rules, where an ontology coupled with SWRL rules for annotating pseudogenes and answering research questions has been proposed in (Holford *et al*., 2010). All the above papers present the need for extending ontologies with rules in order capture the knowledge of complex biomedical domains.

The paper presents Panacea, a semantic-enabled system for discovering drug recommendations and interactions. Panacea is based on experiences and lessons drawn from the development of GalenOWL (Doulaverakis *et al*., 2012), a similar system which had Semantic Web technologies in its core. As such, Panacea can be considered the evolution of GalenOWL in terms of design and scalability. Panacea makes use of established and standardized medical terminologies together with a rich knowledge base of drug-drug and drug-diseases interactions expressed as rules. Panacea is implemented having in mind scalability, completeness of results and responsiveness in query answering.

The paper is organized as follows: Section 2 gives details about the architecture and usage of the framework. In Section 3 the data modelling approach is presented and in Section 4 the core ontology and the layered reasoning process is described while also outlying

*To whom correspondence should be addressed: doulaver@iti.gr

two approaches for rule-based reasoning. Section 5 gives an evaluation of the framework it terms of scalability ans performance and the paper concludes with Section 6.

2 ARCHITECTURE AND FUNCTIONAL DESIGN

The purpose of Panacea is to provide drug prescription recommendations based on a patient's medical record, i.e. advise physicians to prescribe medications according to the drugs active substance indications and contraindications. For details regarding the initiative that triggered development of Panacea and the initial medical and pharmaceutical data that were available, the reader is encouraged to read (Doulaverakis *et al*., 2012). Panacea has been developed in the Java programming language and is built using Apache Jena. Jena provides the API and methods to translate the medical knowledge of terminologies and pharmaceutical rules to semantic entities, while also providing the reasoning engine to enrich the knowledge base.

Panacea follows a layered reasoning process which is depicted in Figure 1. During the start-up of the system, the medical terminologies, namely ATC, UNII, ICD -10 and custom encodings, are transformed to semantic entities, using an appropriate vocabulary, and the initial ontology is constructed. The ontology binds to a reasoner to infer relations such as inheritance and unions. This process is performed once offline during initialization and the knowledge base is available to the system for further utilization. In order to get recommendations in Panacea, a patient instance with the appropriate medical record data is created and fed to the knowledge base. The reasoning process enriches the patient instance with inferred knowledge, thus making it explicit. On this enriched instance, and by utilizing a different reasoning process, the set of medical rules is applied upon. The result of this final stage of rule-based reasoning is the recommendations list which can be retrieved through SPARQL querying.

A key characteristic of the suggested architecture is that, regarding second level reasoning, the framework can utilize any rule-based reasoner or rule engine. Since all the inferred knowledge of the medical definitions and patient data is materialized in the knowledge base, the medical rules can be expressed and loaded in an appropriate rule engine. The rule engine could be an ontology reasoner, such as Jena's reasoning engine, or a business rule manager such as Drools or even CLIPS with appropriate customizations in the data structures. This approach helps in bringing together the best of both worlds: semantic and meaningful representation of data using Semantic Web technologies and the maturity of traditional rule engines in efficiently handling complex and large amounts of rules.

2.1 Use case scenario

In order to demonstrate the benefits of the proposed semantic recommendation system, a use case regarding a possible scenario is described: An elder man visits his family doctor complaining for pain in his right lower back and abdominal region which is accompanied with fever. After appropriate clinical examination, he is diagnosed with right pyelonephritis (ICD-10 code: N11.0). According to the patient's medical history, he is suffering from chronic atrial fibrillation for which he receives clopidogrel (ATC code: B01AC04), vertigo for which he receives cinnarizine (ATC: N07CA02), high arterial blood pressure for which he receives candesartan (ATC: C09CA06) and amlodipine (ATC: C08CA01), and

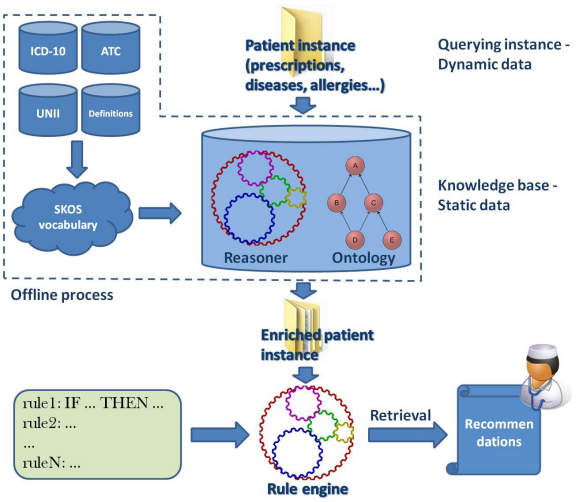

Fig. 1. Panacea framework architecture and data flow

diabetes mellitus for which he receives metformin (ATC: A10BA02) and sitagliptin (ATC: A10BH01). For the new condition of pyelonephritis that was diagnosed, the treating doctor must decide a number of things. Regarding the prescription for treating this new disease, the doctor has to decide which active substances to prescribe in order to treat the resulting inflammation, the cause of the inflammation, the back and abdominal pain and the resulting fever. However, before a decision is made the following factors regarding the patient's medical history should also be considered:

- There should be a check for drug-drug interaction that the patient is taking, before the onset of the new condition (the pyelonephritis)
- There should be a check for drug-disease interaction that the patient is taking, with the new condition
- The new prescription has to be verified that it will not have adverse effects or interactions with the previously prescribed medication and with the patient's medical history

It is clear that the task for the doctor can be hard and a misjudgement could lead to wrong prescriptions. Using an automated drug recommendation system can minimize this risk. The recommendation system will use the input data and the pharmaceutical rules in order to infer a treatment that will be safe for the patient.

3 SEMANTIC TRANSFORMATIONS

Panacea is built on top of international standards of medical terminology in order to represent medical and pharmaceutical information. The following standard terminologies are used:

ICD-10: The World Health Organization classification of diseases. It is used in Panacea for unique identification of diseases thus uniquely identifying drug indications and contraindications related to diseases.

UNII: Unique Ingredient Identifier. Used for the identification of active ingredients found in drugs. In Panacea it is used for uniquely identifying drug indications and contraindications related to ingredients.

ATC: The Anatomical Therapeutic Chemical Classification is used for the classification of drugs. In Panacea it is used in similar fashion to UNII.

IVT: The International Virus Taxonomy is used for the classification of viruses. In Panacea it is used in order to uniquely drug indications and contraindications related to viruses.

Apart from these international standards, a number of domain classifications have been declared and used in order to enhance the usability of the system or to represent data that are not included in the standards. These classifications act as supplementary to the standards.

Substance: As the use of encodings for drug ingredients is not convenient for humans, the identification of active substances is done using its common name references in medical bibliography. These names come from international standards such as the International Nonproprietary Names (INN) and others such as USAN (United States Adopted Name) or BAN (British Approved Name). Members of this identification list are substances such as *acetazolamide* or *isradipine*. In addition, substances correspond to ATC codes such that for example *acetazolamide* $\equiv$ S01EC01. The substances are the actual recommendations of Panacea.

Custom Concepts: While the ATC, ICD-10, UNII and IVT standards are complete, they are designed for use in contexts different from Panacea and drug recommendations, e.g. for annotation, search or information retrieval. As such, it is often desirable to enrich the knowledge base with information that, while not standard, will aid in the usability and overall efficiency of the system. Especially for medical/pharmaceutical rules formulation, it was found out that there were occasions that the definition of diseases, drugs or other was either absent, incomplete or too general to be useful for a rule definition. An example for the lack of a definition in ICD-10 is the absence of a precise and specific code for "Chronic obstructive pulmonary disease" or for "Hypertrophy (benign) of prostate". For this reason, a number of custom concepts have been defined. Examples of such concepts is disease definition such as "Narcolepsy", microorganisms such as "clostridium clostridiiformis" or medical acts such as "upper extremity arteriography".

Custom Concept Collections: Certain "groups" of substances and/or diseases are frequently present in drug interactions and these groups are not recorded explicitly in any standardized classification, so it's more convenient for medical use to specify these custom groups. These often used groups are termed "conditions" in Panacea and are defined by medical experts. A condition can appear as a premise in other condition definitions, as in the Custom Concept Collection *cardiac-rhythm-abnormalities* below, thus enabling their recursive definition:

cardiac-rhythm-abnormalities = cc:bradycardia | icd:R00 | cc:tachycardia | icd:O68.0 | icd:O68.2

where *cc:bradycardia* is defined as (*icd:I49.5* | *icd:R00.1* | *icd:O68.0*) and *cc:tachycardia* as (*icd:R00.0* | *icd:I49.5* | *icd:I47* | *icd:O68.0*). "*icd:*" stands for the ICD-10 namespace.

3.1 SKOS vocabulary

In the approach followed in (Doulaverakis *et al*., 2012), the medical standards and the custom definitions were translated to OWL classes, primitive and defined. While this approach had the benefit of using the language's semantics to model the available information, there were problems resulting from this design decision. One of the major issues was the difficulty in scaling the system. Until currently, very few reasoners are available that can efficiently handle the amount of class definitions and reasoning required to run the system, both in terms of memory consumption and speed.

In Panacea, a different approach was adopted. The SKOS (SKOS: Simple Knowledge Organization System, 2009) vocabulary is a W3C recommendation, it's built using RDF(S) semantics and has been developed as a low-cost migration path for porting existing knowledge organization systems, such as thesauri, taxonomies, classification schemes and subject heading systems, to the Semantic Web. It enables a "lightweight" semantic representation of such knowledge systems and is a good match for the medical standards that are used in Panacea. As such, all the terminologies which are mentioned in the previous section have been transformed using the SKOS vocabulary automatically using a parser.

Comparing SKOS to the approach followed in (Doulaverakis *et al*., 2012), instead of representing the ATC, ICD-10 and UNII classifications as top-level classes, they are now represented as instances of the *skos:ConceptScheme* class. "*skos:*" stands for the SKOS namespace. Each entry in these classifications is represented as an instance of the *skos:Concept* class. The OWL class hierarchy of (Doulaverakis *et al*., 2012) is represented in Panacea using the properties *skos:broaderTransitive* and *skos:narrowerTransitive*, while the unions of classes for Custom Concepts Collections are represented using the *skos:member* property. Correspondence between the semantic transformation methodologies that were followed in the current work and in (Doulaverakis *et al*., 2012) is presented in Table 1.

It is interesting to note that the SKOS vocabulary offers exactly what is needed in order to capture the semantics of the medical classifications without making sacrifices in expressiveness. One can argue that it can be considered more precise than the OWL expressions, as in the case of the similarity of Substances and ATC codes. This similarity is better represented by the *skos:closeMatch* relation than *owl:equivalentClass*. For Panacea a total of $64,658$ definitions of classification codes have been expressed using SKOS.

4 PANACEA ONTOLOGY AND REASONING

The core ontology of Panacea is visualized in Figure 2. The aforementioned SKOS ontologies were imported to the Panacea core ontology under the *MedicalDefinitions* class. The Patient class holds the patient instances and is connected to the *MedicalDefinitions* class with the *hasData* properties. The patient recommendations, indications and contraindications, regarding substances that should and should not be prescribed are expressed with the *canTake* and *cannotTake* properties, respectively. The patients age group and sex group are expressed through the *hasAgeGroup* and *hasSexGroup* properties.

4.1 Medical reasoning

When querying the system for recommendations, a patient instance is created with the initial patient data (through the *hasData*, *hasAgeGroup* and *hasSexGroup* properties) and is loaded in the knowledge base. The reasoner, using RDFS inference and a small number of additional rules, infers all the implicit patient data. As an example consider a patient who suffers from

Table 1. Correspondence between the semantic transformation in the early GalenOWL system and the proposed Panacea framework

	GalenOWL	Panacea
Annotation	rdfs:label	skos:prefLabel
Equivalence	owl:equivalentClass	skos:closeMatch
Custom collections	owl:unionOf	skos:member
Hierarchy	rdfs:subClassOf	skos:broaderTransitive

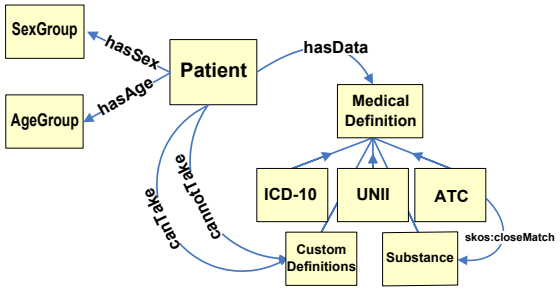

Fig. 2. Panacea ontology

a form of thrombocytopenia. An instance is created with the property *<pan:patient pan:hasData icd:D69.6>*. The reasoner through the skos:broaderTransitive relation will infer the triples *<pan:patient pan:hasData icd:D69>, <pan:patient pan:hasData icd:D65-D69>, <pan:patient pan:hasData icd:D50-D89 >*. Additionally, the custom collection definition of *pnc-cc:deficiency-bone-marrow* has *icd:D69.6* as one of its members so the triplet *<pan:patient pan:hasData pnc- cc:deficiency-bone-marrow>* will also be inferred. At the end, the patient instance will be enriched with all the underlying implicit information.

4.2 Rule-based reasoning

Drug recommendations in Panacea are generated using a rule-based approach. The rules express the indications and contraindications of drug substances while their premises are the medical definitions and the patients' age and sex group. The rules use the logical operators *and* (&) and *or* (|) and parentheses. An example of a rule is for the substance " lisuride" which is expressed as

lisuride = icd:E22.0 | (icd:E22.1 & (icd:N91.0 | icd:N97)), ageGroup=*adult* or *elder*

The above rule reads that the substance "lisuride" is recommended for adult and elder patients who suffer from E22.0, OR suffer from E22.1 AND one of the N91.0 OR N97. For using these rules, they have to be properly parsed and transformed in order to match the knowledge base and the enriched, with implicit knowledge, patient instance. The proposed rule structure allows modifications to specific rules without the changes affecting the rest of the rule base. This enables the rule base to be up-to-date with the latest clinical advancements, which is a requirement as clinical pharmacology and medicine are constantly evolving. Analysing Panacea's architecture in Figure 1 it can be seen that due to the layered reasoning approach, the knowledge base (medical definitions + reasoner) is actually used for producing the enriched patient instance. This means that the instance can be fed to a rule reasoner which has appropriately loaded the medical-pharmaceutical rules, without the reasoner having to communicate with the knowledge base for further utilization. Using this approach and with proper modifications, any rule engine can be used to produce the drug recommendations. To demonstrate this ability, two separate rule engine integrations have been developed and are presented below. The medical rule base consists of $1,342$ rules which were extracted and encoded directly from official documents, such as Summary of Product Characteristics (SPC) Patient Information Leaflets (PIL), regarding drug indications, contraindications, interactions and dosage. The validity of the rule base has already been assessed in (Doulaverakis *et al*., 2012).

It should be noted that work is under way in order to add more functionalities in the drug proposed recommendation system. One of these is the ability to offer additional information such as the proposed dosage for a recommended substance. In order to accomplish such a task, the pharmaceutical rules are being enriched with clinical variables that are important, other than sex and age group. These variables include somatometric characteristics such as height and body weight, creatinin clearance (useful for calculating the dosage for antineoplasmic drugs) and the disease itself as a substance could be indicated at a specific dosage to treat a certain disease, but a different dosage is recommended for another disease.

4.2.1 Jena rule engine For using the rule engine of the Apache Jena API (Apache Jena, 2012) the rules had to be translated to the Jena rule language. An automated parser was developed for this purpose. As for most semantic rule reasoners, OR clauses are not allowed in a rule definition so separate rules had to be expressed for every premise that was OR'ed in the original rule base. For example, the rule for "lisuride" was expressed by 3 different rules:

```
1. (?patient pan:hasData icd:E22.0) →
     (?patient pan:canTake sub:lisuride)

2. (?patient pan:hasData icd:E22.1)
   (?patient pan:hasData icd:N91.0) →
     (?patient pan:canTake sub:lisuride)

3. (?patient pan:hasData icd:E22.1)
   (?patient pan:hasData icd:N97) →
     (?patient pan:canTake sub:lisuride)
```

This rule expansion resulted in a total of $6,451$ rules to be expressed in the Jena language. Trying to load the whole rule base and performing inference for recommendations proved inefficient for

real time use, requiring on average as much as 8 seconds. In order to tackle this issue a coarse rule selection phase was introduced. The selection was executed in 2 iterations. During the first iteration, a subset $\mathcal{A}$ of candidate rules is created from the initial rule base, that match the patient's sex and age group. This subset is selected for further processing. In the second iteration, rules from $\mathcal{A}$ that contain at least one of the patient's data, i.e. a *skos* term, in their premises are singled out and a final set $\mathcal{R} \subseteq \mathcal{A}$ is created from them. Remembering that the implicit knowledge extraction was performed during the introduction of the patient instance to the reasoning framework, creation of $\mathcal{R}$ is actually a simple and fast process. It merely requires string matching and all the whole processing is executed in memory. As a result the overall burden that is added to the whole reasoning process is minimal. From the initial rule base of $6,451$ rules it is common for $\mathcal{R}$ to contain as less as 50 rules, whose evaluation is much more efficient. Rule execution is performed with the Jena rule engine and the patient instance is modified and now contains the drug recommendations. These recommendations are retrieved through SPARQL querying, using Jena's query engine. The advantage of the Jena engine is that it can readily consume the patient instance for producing the recommendations.

4.2.2 Drools rule engine As an alternative approach, the Drools (Drools, 2012) business rule engine was used. In contrary to the Jena engine, Drools could not directly use the patient instance for performing reasoning. For this purpose, the instance was transformed to a Java bean, where the properties of the ontology Patient class are mapped to Java methods using the JenaBean API (`http://code.google.com/p/jenabean/`). The bean was appropriately declared to Drools and was handled for rule execution. A similar approach for integrating Jena and Drools was used in (Bragaglia *et al*., 2010). The Drools Rule Language (DRL) permits the use of OR'ed clauses in the body, so the $1,342$ original medical rules were translated to the same amount of rules in DRL, using an automated parser similar to the one used in the Jena approach. For example, the rule for "lisuride" from the previous paragraph was expressed in DRL as:

```
RULE ''lisuride''
 WHEN
 p: Patient(data : hasData)
 exists( (MedicalDef(uri==icd:E22.0)
    from data) ||
 (MedicalDef(uri==icd:E22.1 && uri==icd:N91.0)
    from data) ||
 (MedicalDef(uri==icd:E22.1 && uri==icd:N97)
    from data) )
 THEN
 Substance lisuride = (Substance)JenaBean.
    reader().load(sub:lisuride);
 modify(p) {p.canTake(lisuride)}
 END
```

Execution was straightforward with no preprocessing required. Drools is optimized for handling large rule bases, so no rule preselection step was required as this would have little impact in reasoning efficiency. The result of this reasoning process is a modified patient Java bean with the drug recommendations. The bean is transformed to Jena model instance and SPARQL querying for retrieving the recommendations is possible. What this approach demonstrates is that it's possible to integrate business rule engines as reasoners in the framework, thus being able to make use of the high efficiency and optimizations of these engines with the semantic description and interpretation of data.

5 EVALUATION AND DISCUSSION

For evaluating the framework, a comparison was made between the two approaches for the final stage reasoning and with GalenOWL (with values taken from (Doulaverakis *et al*., 2012)). The comparisons were focused on the usability of the framework in a production environment as the rule base has been validated in (Doulaverakis *et al*., 2012). Three parameters were measured. These were initialization time, the time to get the system up and running, memory consumption after initialization, and query response time, i.e. the time that is needed to have the rule base executed and the results retrieved. Results are shown in Table 2.

There are some points to discuss in the table results. Initialization involves loading the ontology in memory, performing inference, and preparing the medical rule base for patient data reasoning. In the Jena implementation, the rule base is processed and loaded only after the patient instance has been introduced to the system, while the Drools implementation loads the whole rule base on the engine before any patient data are introduced. As a result, Drools appears slower than the Jena approach regarding initialization. For the same reason, memory consumption appears greater for Drools. This metric corresponds to memory consumption from initialization to recommendations retrieval. While in Drools the whole rule base is loaded on memory, in Jena the approach was to load a small subset of the rule base that could possibly match the patient data, which leads to a smaller memory footprint. Finally, for query response the advantage is with Drools, as was expected, mainly due to the fact that Drools is a dedicated rule engine while Jena's focus is not at providing a state of the art reasoner and rule engine, but a versatile API for ontology management.

Numerically, the Jena approach seems to be more efficient than Drools, apart from the query execution time but for which the difference is not important. However, while for the present knowledge base Jena seems to perform better, this fact could change as more and more rules are added. It is estimated that eventually at its final stage, Panacea will incorporate more than $9,000$ drug-drug and drug-disease interactions. As already said, Jena is more focused as an ontology API and less as an efficient rule engine which could eventually lead to scaling problems. On the other hand, scaling with Drools is not an issue. The value of business rule engines as Semantic Web reasoners has been previously exploited using approaches such as (O'Connor and Das, 2012), where the authors implemented two OWL2-RL Motik *et al*. (2009) reasoners using the Drools and Jess rule engines respectively. The use of traditional rule engines with the Semantic Web technologies brings together the best of both worlds, i.e. increased efficiency coupled with interoperability and semantic annotation of information.

What is also noticeable from Table 2 is the decreased memory requirement of Panacea compared to the previous OWL- based GalenOWL system, although the two approaches offer very similar functionality. As a result of this achievement, Panacea can accommodate a far greater knowledge base thus supporting the claim of increased scalability.

Panacea will eventually be offered as a service with potential customers being health care professionals. Other possible exploitation routes are being investigated such as integration to patient

Table 2. Evaluation between the 2 Panacea reasoning approaches and GalenOWL

	Panacea-Jena	Panacea-Drools	GalenOWL
Initialization time	32.0 s	34.7 s	148 s
Memory consumption	169 MB	280 MB	649 MB
of which rule base consumes	0 MB	111 MB	—
Query response time	47 ms	5 ms	16 ms

management systems in health clinics. The use of personalized drug prescription systems, as Panacea, in everyday practice will have major advantages to the society and the economy. A major benefit from the use of such systems is the reduction of medical costs through rational drug prescriptions that personalized drug prescription allows (Fischer *et al*., 2008). Another benefit is a positive effect in public health with reduction of outbreaks relating to drug interactions or adverse effects (Ammenwerth *et al*., 2008). All knowledge regarding drug information is encoded and is available to the experts in order to aid them during prescriptions thus acting as decision support systems. It should be stressed out that drug recommendation systems do not aim to replace medical experts but to support them in their practice.

A limitation of the proposed approach is that a rather large amount of manual effort by experts is required in order to populate and enrich the rule base. Although the semantic technologies that have been employed can make rule authoring simpler, no automated method for pharmaceutical rule generation has been integrated. However, one would argue that since rule authoring is performed by experts then the rules are verified and guaranteed to be correct. Even if an automated method, such as rule mining, had been implemented, the generated rules would still have to be verified be an expert in the field. Manual verification, although less intensive, would still be required.

6 CONCLUSIONS

The paper presented Panacea, a framework for semantic-enabled drug recommendations discovery. The framework utilizes a layered reasoning approach were the medical ontology and the patient data instances are fed to an extended RDF(S) reasoner in order to infer implicit knowledge. Drug recommendations are generated using the second reasoning layer where any common rule engine can be used. As a proof of concept implementation, the Jena reasoner and the Drools rule engine has been integrated and separate tests regarding requirements and efficiency were conducted. For the Jena reasoner implementation a 2 step rule selection method was followed which resulted in computationally efficient reasoning. The tests concluded that both approaches perform significantly better than the earlier GalenOWL system, while at the same time maintaining the same quality of results and improving performance. Concerning future work, the addition of dosage recommendations in the rules is an ongoing work. Additionally, the possibility to add probabilistic reasoning will be investigated. To this end, Drools is being extended with a fuzzy reasoning engine (Sottara *et al*., 2008), which while it's still in development, it's actively supported and it is mature enough to be able to use it as a testing framework.

ACKNOWLEDGMENTS

This work has been supported by the national project "Panacea", funded by GSRT Hellas under the "Support for SMEs" programme.

REFERENCES

Adnan, M., Warren, J., and Orr, M. (2010). Ontology based semantic recommendations for discharge summary medication information for patients. In *Computer-Based Medical Systems (CBMS), 2010 IEEE 23rd International Symposium on*, pages 456–461.

Ammenwerth, E., Schnell-Inderst, P., Machan, C., and Siebert, U. (2008). The effect of electronic prescribing on medication errors and adverse drug events: A systematic review. *J Am Med Inform Assoc*., **15**(5), 585–600.

Apache Jena (2012). Semantic web framework. http://jena.apache.org/.

Beuscart, R., McNair, P., J, J. B., and PSIP consortium (2009). Patient safety through intelligent procedures in medication: the PSIP project. *Studies in Health Technology and Informatics*, **148**, 6–13.

Bragaglia, S., Chesani, F., Ciampolini, A., Mello, P., Montali, M., and Sottara, D. (2010). An hybrid architecture integrating forward rules with fuzzy ontological reasoning. In *Proceedings of the 5th international conference on Hybrid Artificial Intelligence Systems - Volume Part I*, HAIS'10, pages 438–445, Berlin, Heidelberg. Springer-Verlag.

Doulaverakis, C., Nikolaidis, G., Kleontas, A., and Kompatsiaris, I. (2012). GalenOWL: Ontology based drug recommendations discovery. *Journal of Biomedical Semantics*, **3**(14).

Drools (2012). Business logic integration platform. http://www.jboss.org/drools.

Fischer, M., Vogeli, C., Stedman, M., Ferris, T., Brookhart, M., and Weissman, J. (2008). Effect of electronic prescribing with formulary decision support on medication use and cost. *Archives of Internal Medicine*, **168**(22), 2433–2439.

Golbreich, C. (2004). Combining rule and ontology reasoners for the semantic web, invited talk, rules and rule markup languages for the semantic web. In *Boley Editors, LNCS 3323*. Springer.

Golbreich, C., Dameron, O., Bierlaire, O., and Gibaud, B. (2005). What reasoning support for ontology and rules? the brain anatomy case study. In *Proceedings of the KR 2004 Workshop on Formal Biomedical*, pages 60–71.

Holford, M., Khurana, E., Cheung, K.-H., and Gerstein, M. (2010). Using semantic web rules to reason on an ontology of pseudogenes. *Bioinformatics [ISMB]*, **26**(12), 71–78.

Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., and Lutz, C. (2009). OWL 2 Web Ontology Language Profiles, W3C recommendation. http://www.w3.org/TR/owl2-profiles/.

O'Connor, M. and Das, A. (2012). A Pair of OWL 2 RL Reasoners. In *Proceedings of OWL: Experiences and Directions Workshop 2012 (OWLED-2012)*, Heraklion, Greece.

Sheth, A. (2005). Semantic web & semantic web services: Applications in healthcare and scientific research, keynote talk. In *IFIP Working Conference on Industrial Applications of Semantic Web*, Jyvaskyla, Finland.

SKOS: Simple Knowledge Organization System (2009). W3C recommendation. http://www.w3.org/2009/08/skos-reference/skos.html.

Sottara, D., Mello, P., and Proctor, M. (2008). Adding uncertainty to a rete-oo inference engine. In *Proc. of the International Symposium on Rule Representation, Interchange and Reasoning on the Web. RuleML '08*, pages 104–118.

Ontology modeling of genetic susceptibility to adverse events following vaccination

Yu Lin, Yongqun He

Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, Center for Computational Medicine and Bioinformatics, and Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109, USA

ABSTRACT

Administration of different vaccines triggers a variety of adverse events in some groups of people but not in others. This phenomenon may be due to the variation of genetic factors that affects the susceptibility to vaccine adverse events. In this study, we introduce the development of an Ontology of Genetic Susceptibility Factor (OGSF) that is aligned with the Basic Formal Ontology (BFO). OGSF represents the genetic susceptibility, genetic susceptibility factors and vaccine adverse events using formal ontologies. Two case studies were used to test and validate the model. One case study represents a human gene allele DBR1*15:01 as a genetic susceptibility factor to vaccine Pandemrix related multiple sclerosis. Genetic polymorphisms associated with smallpox vaccine adverse event was analysed as the second use case. A SPARQL query, visualization of extracted data as a network and the social network analysis of the network, further provide insights on the evaluation and application of the ontology.

1 INTRODUCTION

Vaccines have enabled tremendous decreases in infectious diseases and remain among the most effective of our public health initiatives. At the same time, as an ever increasing number of vaccines is administered globally, many vaccine-associated adverse events and reactions have been identified and threaten the public health successes attributable to vaccines [1]. As defined in the Vaccine Adverse Event Reporting System (VAERS) and Ontology for Adverse Event (OAE), a vaccine adverse event is an adverse event following vaccination and does not assume a causal association [2]. Vaccine-related adverse events often occur in some populations but not in others, which has led to the hypothesis of genetic susceptibility to vaccine adverse events [3, 4].

Genetic susceptibility, also called genetic predisposition, is an increased likelihood or chance of developing a particular disease due to the presence of one or more gene mutations and/or a family history that indicates an increased risk of the disease. The allele that confers the increased risk/susceptibility may be inherited but the disease itself will not. The single locus genotype is usually insufficient to cause a disease. For the disease to appear, impaired expressions of alleles at other gene loci and/or environmental factors are often needed [5].

Genetic susceptibility factors are the genetic entities, most likely genetic variations, which influence the susceptibility. The genetic susceptibility factors contributing susceptibility to a disease may not be obvious mutations. It is more likely a combination of subtle changes on several genes, which may be quite common in the healthy population. Moreover, the main determinants of susceptibility may be different in different populations [6]. With current technological advances and new biostatistics approaches to understanding a large number of databases of information, we can now better understand how genetic variations become critical to vaccine-induced positive host responses and adverse reactions.

An Ontology of Genetic Susceptibility Factor (OGSF) was previously developed for our formalization of the definitions of 'genetic susceptibility' and 'genetic susceptibility factor' using the TCF7L2 gene and its susceptibility to Type 2 Diabetes as an example [7]. The entities important for the representation of genetic susceptibility to diseases include: genetic polymorphism, the population and geographical location, the disease entities, and related statistical entities (e.g., odds ratio and p-value). Here we consider that a vaccine adverse event is a pathological bodily process, and we extend the former work to model the genetic susceptibility to adverse event.

Based on previous studies, we have now developed a new version of genetic susceptibility-focused ontology, the Ontology of Genetic Susceptibility Factor (OGSF) by using Basic Formal Ontology (BFO) 2.0 as its upper ontology. OGSF is used to study the susceptibility factors associated with vaccine adverse events.

2 METHODS

2.1 Ontology editing

The format of OGSF ontology is W3C standard Web Ontology Language (OWL2) (http://www.w3.org/TR/owl-guide/). For this study, many new terms and logical definition were added into original OGSF [7] using the Protégé 4.3.0 build 304 OWL ontology editor (http://protege.stanford.edu/).

2.2 Ontology term reuse and new term generation

OGSF imports the whole set of the Basic Formal Ontology (BFO) [8]. To support ontology interoperability, many terms from reliable ontologies are reused. For this purpose, OntoFox [9] was applied for extracting individual terms from external ontologies. For those genetic susceptibility-specific terms, we generated new OGSF IDs with the prefix of "OGSF_" followed by seven-digit auto-incremental digital numbers.

* To whom correspondence should be addressed: yuln@med.umich.edu

2.3 Evaluation of OGSF

Use case studies were designed based on literature survey. SPARQL was performed using the SPARQLquery plug-in embedded with Protégé 4.3.0 build 304. Graphed data was extracted using the OntoGraf plug-in [10] Gephi 0.8.2 beta (http://gephi.org)[11] was used to conduct social network data analysis and visualization based on the extracted graph data from instances of OGSF.

2.4 Availability and access

The website for OGSF project is available at http://code.google.com/p/ogsf/. The source of the ontology is also available in the NCBO Bioportal: http://bioportal.bioontology.org/ontologies/3214.

3 RESULTS

3.1 OGSF is aligned with BFO

The development of OGSF follows the OBO Foundry principles, including openness, collaboration, and use of a common shared syntax [12]. The early version of OGSF was not well aligned with BFO. To align OGSF with BFO 2.0 Graz version, we started with key terms and render them using BFO's terms as parent terms.

There are two core terms in OGSF: 'genetic susceptibility' and 'genetic susceptibility factor'. The OGSF term 'genetic susceptibility' (OGSF_0000000) is a subclass of 'disposition' (BFO_0000016). The alternative term for 'genetic susceptibility' is 'genetic predisposition'. Note that in BFO 2.0 the term 'predisposition' is not included, so we put genetic susceptibility directly as the child term of 'disposition'. The first level child terms of 'genetic susceptibility' include: 'genetic predisposition to disease of type X' (OGMS_0000033), 'genetic susceptibility to pathological bodily process' (OGSF_0000001), and 'genetic susceptibility to biological process' (OGSF_0000002). The term that reveals our use case is 'genetic susceptibility to adverse event following vaccination' (OGSF_0000010) and it is the third level child term of 'genetic susceptibility'.

Another core OGSF term 'genetic susceptibility factor' (OGSF_0000004) is a subclass of 'material entity' (BFO_0000040). An allele, gene, genotype, and haplotype can be genetic susceptibility factors. The relation: 'material basis of at some time' (BFO_0000127), is used to link genetic susceptibility factor and genetic susceptibility.

3.2 Modeling genetic susceptibility to adverse event following vaccination

The genetic susceptibility to vaccine adverse events is used as a use case for OGSF redesign.

Genetic susceptibility reflects the relation between a genetic factor (e.g. allele) and risk of condition, disease or responses to vaccines or drugs. Different levels of genetic association studies, such as family studies, genetic linkage studies, and population-based studies are conducted in order to determine whether or not a genetic variation mediates the diseases outcome such as a vaccine adverse event.

Fig. 1 shows how we use OGSF terms and BFO relations to represent genetic susceptibility to vaccine adverse event.

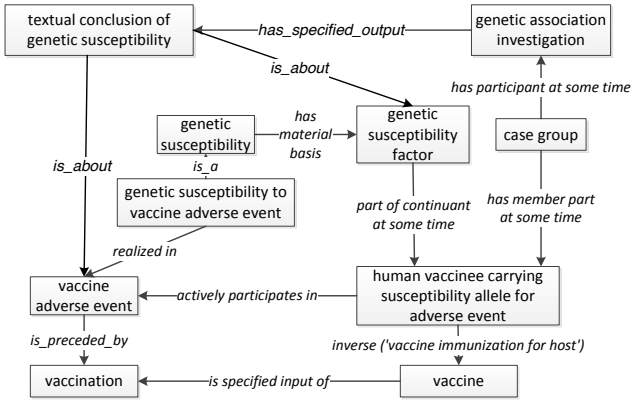

Fig. 1. Design pattern for representing genetic susceptibility to a vaccine adverse event (VAE).

The set of core terms representing the whole topic are 'genetic susceptibility factor', 'genetic susceptibility', 'adverse event' and 'textual conclusion of genetic susceptibility'. In Fig.1, the 'genetic susceptibility factor' is the material basis of 'genetic susceptibility', which has a subclass 'genetic susceptibility to vaccine adverse event'. The genetic susceptibility is realized in the process of 'vaccine adverse event'. The 'genetic susceptibility factor' is the part of a 'human vaccinee carrying susceptibility allele for adverse event', which 'actively participates in' the 'vaccine adverse event'. On the other hand, a 'genetic association investigation' has participant 'case group' with the 'human vaccine carrying susceptibility allele for adverse event' as its member. The 'genetic association investigation' has 'textual conclusion of genetic susceptibility' as its specified output, and the conclusion 'is about' both 'genetic susceptibility factor' and 'vaccine adverse event'. An inverse of VO relation: 'vaccine immunization for host' interlinks the human vaccinee and 'vaccine'. 'Vaccine' is a specified input of the process of 'vaccination'. Relation 'is preceded by' linking 'vaccination' and vaccine adverse event' indicates that 'vaccination' happens before the 'vaccine adverse event'.

3.3 Modeling genetic association study

Studies have provided many supporting evidences for asserting susceptibility factors to adverse event outcomes. Based on the OBI framework, we specially modeled the genetic association study designs according to our use case. The textual definition of OGSF term 'genetic association investigation' was given as: 'an investigation that aims to test whether single-locus alleles or genotype frequencies (or more generally, multilocus haplotype frequencies) differ

between two groups of individuals (usually diseased subjects and healthy controls)'. Different types of those studies exist, such as 'case-control study', 'GWAS study' (Genome-Wide Association Study) and 'case report'. 'GWAS study' is a type of 'case-control study' and has two subclasses 'initial GWAS study' and 'replicate GWAS study'. The statistical method conducted in a study is modeled as 'data analysis' that is a part of an investigation as asserted in OBI. 'Case group' and 'control group' are subclasses of 'human study subject group'. The 'human study subject group' is the participant of the 'genetic association investigation'.

A statistical analysis of the genetic susceptibility is based on the choice of a statistical study design, which depends on several factors related to the phenotype: the population, the accurate measurement of environmental factors, and known genetic background among other factors. Due to the presence of many different cofounders, it is often difficult to detect and verify genetic susceptibility factors associated with specific adverse event outcomes. Observed statistically significant genetic susceptibilities may be contradictory among different studies [13]. More and consistent observations in different populations may give stronger evidence to support the true causal relation between a 'genetic susceptibility factor' and an observed outcome. Well-designed experiments may be applied to verify the association. In order to store the result from genetic association studies, we use 'textual conclusion of genetic susceptibility' to be asserted as 'specified output of' a 'genetic association investigation'. The 'textual conclusion of genetic susceptibility' is a 'textual entity'. The 'is about' relation was used to link the conclusion with 1) 'genetic susceptibility factor' and 2) ‘vaccine adverse event’ process.

Three terms: 'positive conclusion of genetic susceptibility', 'negative conclusion of genetic susceptibility' and 'neutral conclusion of genetic susceptibility' are asserted as subclasses of 'textual conclusion of genetic susceptibility'. A 'positive conclusion of genetic susceptibility' means that a positive conclusion is drawn based on a significant statistical association of a genetic factor and a vaccine adverse event as studied in this paper. A 'negative conclusion of genetic susceptibility' a denied association between a genetic factor and an adverse event. Sometimes, depending on the data, an investigator may draw a conclusion of a non-significant association but without a clear deny of a possible association. This situation is captured using ‘neutral conclusion of genetic susceptibility’.

3.4 Case study

Case studies are used for two purposes: 1) to validate the modeling, 2) to test possible applications of the ontology.

3.4.1 Case study 1: HLA allele DBR1*15:01 is genetic susceptibility to Pandemrix related multiple sclerosis

Vrethem *et al.* reported the occurrence of severe narcolepsy with cataplexy and multiple sclerosis (MS) in a previously healthy young male in association with Pandemrix vaccination [14]. The investigators found that that those patients carrying HLA allele DBR1*15:01 were associated with MS and those having HLA allele DQB1*06:02 were associated with narcolepsy. It was also concluded that the genetic susceptibility in this patient is a clue that an immune-mediated mechanism and a common etiology for both diseases in this patient.

The DBR1*15:01 as a genetic susceptibility factor responsible for Pandemrix-induced MS was modeled in the class level using OGSF, and the particular study was modeled in instance level using OGSF (Fig 2).

At the class level, 'DBR1*15:01' is an 'allele of HLA gene', which is also the material basis of (BFO 2: 'material basis of at some time') 'genetic susceptibility to vaccine adverse event'. The instance of ‘DBR1*15:01’ is a part of the MS patient instance. In class level, 'multiple sclerosis AE patient' 'actively participates in' the 'multiple sclerosis AE' process. Multiple Sclerosis adverse event is preceded by the 'Pandemrix vaccination'. 'Pandemrix' is a participant of 'Pandemrix vaccination' and it is related to the MS patient using a short relation from Vaccine Ontology (VO): 'vaccine immunization for host', which relates a vaccine with a vaccinee.

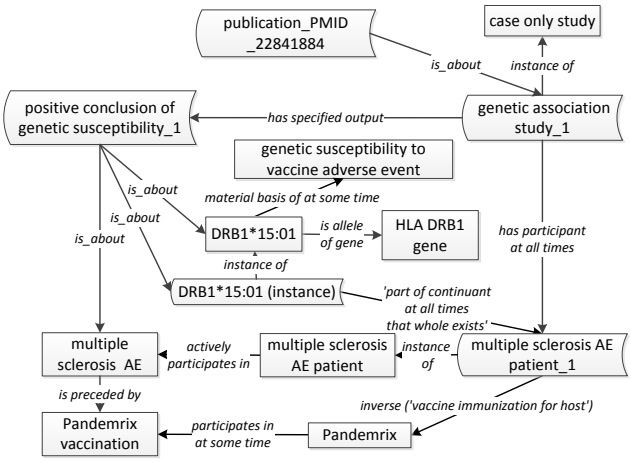

Fig. 2. OGSF modeling of vaccine-associated multiple sclerosis

Since it is a case report, this study gives one specific positive supporting evidence to the genetic susceptibility of DBR1*15:01, which is asserted at the instance level. We use 'genetic association study_1' to represent the study, which gives a specific output 'positive conclusion of genetic susceptibility_1'. This specific conclusion is about the entity 'DBR1*15:01' and the 'multiple sclerosis AE'.

3.4.2 Case study 2: genetic polymorphisms associated with adverse events after smallpox vaccination

Reif e*t al.* reported that genetic polymorphisms in an enzyme methylenetetrahydrofolate reductase (MTHFR) and an immunological transcription factor (IRF1) were associated with AEs after smallpox vaccination [15]. In this study, two independent clinical trials were conducted as initial and replicating genetic association studies separately. The Odds Ratio was used to measure the association between genotypes and systematic adverse event. Only strong association supported by a statistically significant Odds Ratio in both studies was considered and asserted as a true positive genetic association.

In this case, the important information to be stored is the susceptibility allele of the SNPs and the statistical power in two studies. Those information was curated and summarized in Table 1.

Table 1. Statistical summary of genetic susceptibility factors with systematic adverse event following smallpox vaccination

GSF[&]	Allele	Gene	Odds Ratio	P-value	Study 1 or 2
rs1801133 SNP	T	MTHFR	2.3 (1.1–5.2)	0.04	1
rs1801133 SNP	T	MTHFR	4.1 (1.4–11.4)	0.01	2
rs9282763 SNP	G	IRF1	3.2 (1.1–9.8)	0.03	1
rs9282763 SNP	G	IRF1	3.0 (1.1–8.3)	0.03	2
rs839 SNP	A	IRF1	3.2 (1.1–9.8)	0.03	1
rs839 SNP	A	IRF1	3.0 (1.1–8.3)	0.03	2
Haplotype 1[*]	G,A	IRF1	3.2 (1.0–10.2)	0.03	1
Haplotype 1[*]	G,A	IRF1	3.0 (1.0–9.0)	0.03	2
Haplotype 2[#]	T,C,A	IL4	2.4 (1.0–5.7)	0.05	1
Haplotype 2[#]	T,C,A	IL4	3.8 (1.0–14.4)	0.06	2

Notes:
& GSF stands for Genetic Susceptibility Factor
* Haplotype 1 contains G allele of rs9282763, A allele of rs839 in IRF1 gene.
# Haplotype 2 contains T allele of rs2070874,C allele of rs2243268, A allele of rs2243290 in IL4 gene.

The class level assertion is similar to case study 1. For example, the constrains representing one of the genetic susceptibility factors, A allele of rs839, are as follows:

1. *'material basis of at some time' some 'genetic susceptibility to adverse event following vaccination'*
   This axiom denotes that the A allele of rs839 is the material basis of the genetic susceptibility to AE induced by vaccination

2. *'part of continuant at all times that whole exists' some ('human vaccinee experiencing systemic adverse event' and inverse('vaccine immunization for host') some 'Smallpox virus vaccine')*
   This axiom denotes that the 'A allele of rs839' is part of some human who is experiencing systemic adverse event and had vaccinated by Smallpox vaccine

3. *isContainedIn some 'IRF1 gene'*
   This axiom denotes that the 'A allele of rs839' is contained in IRF1 gene

4. *'alternative allele of SNP'*
   This axiom denotes that the 'A allele of rs839' is an alternative allele

5. *'susceptibility allele' (inferred)*
   This axiom denotes that the 'A allele of rs839' is a susceptibility allele, so it is a genetic susceptibility factor.

The instance level representation representing two independent studies provide the statistical supporting evidence to the genetic susceptibility (Fig. 3).

Fig 3 illustrated that two 'positive conclusions of genetic susceptibility' from clinical trail 1 and trail 2 support the 'T allele of rs1801133 SNP' as the 'material basis of at some time' the 'genetic susceptibility of adverse event following vaccination'. The datatype properties 'hasOddsRatio' and 'hasPvalue' are properties of the 'positive conclusion of the genetic susceptibility'. Using these datatype properties, the real data denotes the statistical power was represented in the ontology.

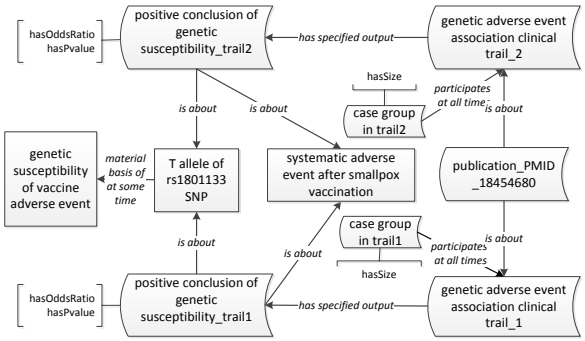

Fig. 3. Modeling Case Study 2 using OGSF

3.4.3 SPARQL query

A SPARQL script was developed to query against inferred OGSF ontology. The query led to the retrieval of the genetic susceptibility factors, as shown in Table 1. (Sparql query script shown in Supplemental material if allowed).

3.4.4 Visualization and social network analysis

In order to give a better view of the terms and links between terms, data from case study 2 was extracted using OntoGraf and visualized using Gephi as following (Fig. 4 and 5).

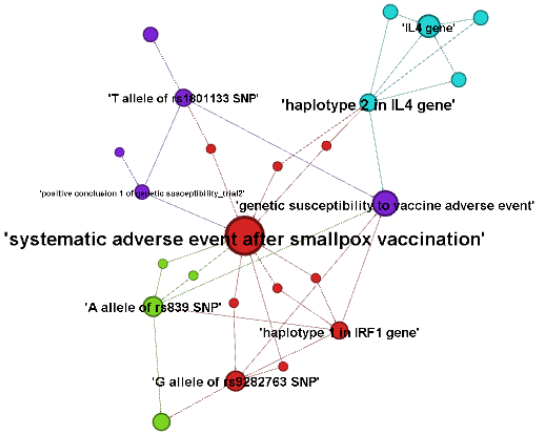

Fig. 4. All related nodes within case study 2.

Fig. 4. shows how the data and terms interlinked with each other in the network of case study2. The most

connected node is 'systematic adverse event after smallpox vaccine', since there are 10 conclusions related to it as shown in table1. All the genes, relevant SNP alleles and haplotypes are interlinked with each other, and can be captured as a community within the network, which indicated by colors of the node.

Running Gephi's 'filter' function, two different views of the network of case study 2 were yield as shown in Fig 5.

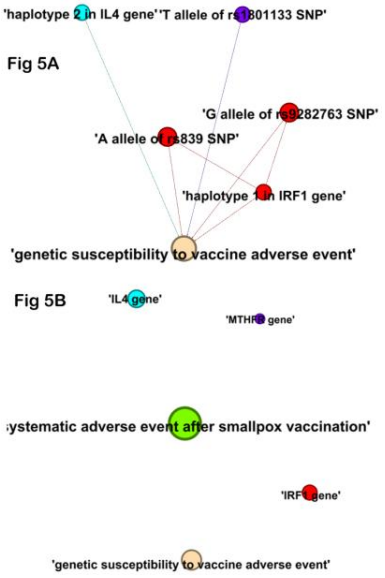

Fig. 5. Two views of the genetic susceptibility network in case study2. (A). EgoNetwork filter view of the network, which shows entities that are directly linked to 'genetic susceptibility to vaccine adverse event'. **(B)** Closeness centrality filtered view of the network. All the dots in the figure have closeness centrality value equal to 0.

Combining Fig. 5A and 5B, it indicates that: 1) in OGSF, the genetic susceptibility is directly related with variants, such as SNPs and haplotypes. 2) Gene is indirectly linked to genetic susceptibility via variants. The in-directed connection can be captured by centrality network analysis in the given data set. In our specific case study 2, the closeness centrality calculations of genetic susceptibility, adverse event and genes are the lowest.

4 DISCUSSION

4.1 Representing genetic susceptibility requires the notion of instance level evidence

The purpose of representing the knowledge of genetic susceptibility here is to extend existing beliefs by adding new facts. For example, if in one study A1, the genetic factor SNP B is statistically significant related to an adverse event C, then the SNP B as a genetic susceptibility factor will be represented using the OGSF framework. This knowledge will become an existing belief, when another study A2 reached the same conclusion, this fact will be added into the OGSF knowledgebase and hence provide stronger supporting evidence to the genetic causal association. Another example is that suppose gene G is related with both SNP B and SNP E, when another study A3 gave the conclusion of SNP E statistically significant related with the same adverse event C. To add this fact into OGSF would strengthen the belief that gene G is related to the genetic causal association.

The notion of genetic susceptibility can be expressed using OWL classes, whereas each study is modeled in instance level as data item. To simplify the connections, the relation 'is_about' was used to bridge the individual level 'textual' conclusions from an individual study to a 'genetic susceptibility factor' (class level) and specific vaccine adverse event (class level). The efficiency and applicable aspects of these relations need to be tested using more complicated datasets and SPARQL query.

4.2 The granularity of genetic susceptibility factor is at allele level

Nowadays, thousands of Single Nucleotide polymorphisms (SNPs) can be tested efficiently in large population-based studies. Researchers are using various entities to describe genetic susceptibility bearers, such as genotype, SNP, LD block, haplotype and so on. Except for LD block, other genetic susceptibility factors can be represented by notion of allele. As defined in our previously developed Ontology for Genetic Interval (OGI) [16], 'allele' is 'an alternative form of a genetic interval that is located at a specific position on a specific chromosome'. In OGI, term 'allele' has following subclasses: 'allele of gene', 'allele of polymorphism', 'allele of SNP', 'allele of phenotype', 'allele shared by sibs'. OGSF fully imports OGI, thus inherited the OGI's allele classes and definitions. OGI gives formalized topological relations between alleles and genes, so that the relations between alleles and genes can be logically calculated [14]. Adopting those relations ensure the example discussed in the section 4.1 can be reasoned in OGSF.

4.3 Visualization of sub network of OGSF data

The ontology's instance level data can be visualized as directed graph. The visualization and network analysis results provide deep insights in terms of ontology designing. Representing the genetic susceptibility can be addressed using three layers of information depending on researchers' interest. The first layer is the direct link of types of genetic factors and investigated adverse event. In our representation, it is grounded to allelic variant level. The second layer is the supporting conclusion that provides positive evidence to the direct link. The third layer is the linking between a gene and the investigated adverse event. Since in OGSF, gene and adverse event are not directly linked, the social network analyses shows that this indirect link can be measured mathematically and thus provide the foundation for prediction. It is noted that usually only significant associations were reported in the literature, and many negative results may not be available. The network analysis may be biased.

In conclusion, based on the formalization of genetic susceptibility, OGSF provides a frame work to represent the genetic allelic variants, genes and pathological processes. It requires ontological scientific discourse representations as those developed in SWAN ontology[17]. Furthermore, a large numbers of databases have been established in order to establish the relation between genotypes and phenotypes. Some of them, such as SNPedia [18], Bio2RDF [19], Leiden Open (source) Variation Database (LOVD) [20] and GWAS central [21], support sematic web and open data technology. OGSF is aim to be an intermediate layer between applications and above existing resources.

ACKNOWLEDGEMENTS

This project was supported by a NIH-NIAID grant (R01AI081062). We would like to acknowledge with appreciation Dr. Wei Zhang, a biostatistician expert from University of Michigan School of Public Health, for his advice and consultation.

REFERENCES

1. Poland GA, Ovsyannikova IG, Jacobson RM: **Adversomics: the emerging field of vaccine adverse event immunogenetics**. *Pediatr Infect Dis J* 2009, **28**(5):431-432.
2. He Y, Xiang Z, Sarntivijai S, Toldo L, Ceusters W: **AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events**. In: *International Conference on Biomedical Ontology: July 26-30, 2011; Buffalo, NY, USA*: http://ceur-ws.org/; 2011: 313-315.
3. Siber GR, Santosham M, Reid GR, Thompson C, Almeido-Hill J, Morell A, deLange G, Ketcham JK, Callahan EH: **Impaired antibody response to Haemophilus influenzae type b polysaccharide and low IgG2 and IgG4 concentrations in Apache children**. *N Engl J Med* 1990, **323**(20):1387-1392.
4. Black FL, Hierholzer W, Woodall JP, Pinhiero F: **Intensified reactions to measles vaccine in unexposed populations of american Indians**. *J Infect Dis* 1971, **124**(3):306-317.
5. Alghabban A: **Dictionary of pharmacovigilance**. London ; Chicago: Pharmaceutical Press; 2004.
6. Strachan T, Read AP: **Human molecular genetics 3**, 3rd edn. New York ; London: Garland Science; 2004.
7. Lin Y, Sakamoto N: **Ontology driven modeling for the knowledge of genetic susceptibility to disease**. *Kobe J Med Sci* 2009, **55**(3):E53-66.
8. Grenon P: **Spatio-temporality in Basic Formal Ontology**. In: *IFOMIS reports.* Edited by Grenon P. Leipzig: 2003: 89.
9. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y: **OntoFox: web-based support for ontology reuse**. *BMC Res Notes* 2010, **3**:175.
10. [http://protegewiki.stanford.edu/wiki/OntoGraf]
11. Bastian M, Heymann S, Jacomy M: **Gephi: an open source software for exploring and manipulating networks.** International AAAI Conference on Weblogs and Social Media, 2009.
12. 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):1251-1255.
13. Bellamy R: **Susceptibility to infectious diseases : the importance of host genetics**. Cambridge, UK ; New York, NY, USA: Cambridge University Press; 2004.
14. Vrethem M, Malmgren K, Lindh J: **A patient with both narcolepsy and multiple sclerosis in association with Pandemrix vaccination**. *J Neurol Sci* 2012, **321**(1-2):89-91.
15. Reif DM, McKinney BA, Motsinger AA, Chanock SJ, Edwards KM, Rock MT, Moore JH, Crowe JE: **Genetic basis for adverse events after smallpox vaccination**. *J Infect Dis* 2008, **198**(1):16-22.
16. Lin Y, Sakamoto N: **Genome, Gene, Interval and Ontology**. In: *2nd Interdisciplinary Ontology Conference: Feb 28-Mar. 1 2009; Tokyo*: Keio University Press Inc.; 2009: 25-34.
17. Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, Ruttenberg A, Clark T: **The SWAN biomedical discourse ontology.** *J Biomed Inform* 2008, **41**(5):739-751.
18. Cariaso M, Lennon G: **SNPedia: a wiki supporting personal genome annotation, interpretation and analysis**. *Nucleic Acids Res* 2012, **40**(Database issue):D1308-1312.
19. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems**. *J Biomed Inform* 2008, **41**(5):706-716.
20. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT: **LOVD v.2.0: the next generation in gene variant databases**. *Hum Mutat* 2011, **32**(5):557-563.
21. Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, Dash D, Brahmachari SK, Brookes AJ: **HGVbaseG2P: a central genetic association database**. *Nucleic Acids Res* 2009, **37**(Database issue):D797-802.

Auditing Redundant Import in Reuse of a Top Level Ontology for the Drug Discovery Investigations Ontology

Zhe He[1,*], Christopher Ochs[1], Larisa Soldatova[2], Yehoshua Perl[1], Sivaram Arabandi[3] and James Geller[1]

[1]NJIT, Newark, NJ, USA, [2] Brunel University, London, UK, [3]Ontopro LLC, Houston, TX, USA

ABSTRACT

The use of a top-level ontology, e.g. the Basic Formal Ontology (BFO), as a template for a domain ontology is considered a best practice. This saves design efforts and supports multi-disciplinary research. The Drug Discovery Investigations ontology (DDI) for automated drug discovery investigations followed the best practices and imported BFO. However not all BFO classes were used. Quality assurance is an important process in the development of ontologies. One methodology proven to support quality assurance is based on automatic derivation of abstraction networks (ANs) from the original ontologies. An AN of an ontology is a compact secondary network summarizing the ontology. ANs were shown to support the identification of sets of concepts with higher concentrations of errors than control sets. In this paper, an AN is derived for the DDI, based on object properties. The top node of this AN represents a set of 81 classes without any object properties. Nodes of an AN representing many classes tend to indicate modeling errors. Upon reviewing these 81 classes, we discovered that among them are most of the classes imported from BFO, and that most of these classes are irrelevant for DDI. An algorithm for hiding such irrelevant classes from a specified ontology is described. As many as 18 (56%) of the 32 BFO classes represented by the top node of the AN were hidden from DDI by the algorithm. We conclude that ontologies reusing a top-level ontology should employ this AN-based approach.

1 INTRODUCTION

The use of a top-level ontology as a template for a domain ontology is considered a best practice in ontology engineering. A well-developed top-level ontology eases the development of domain ontologies and reduces possible errors and inconsistencies. The Open Biomedical Ontologies (OBO) Foundry (www.obofoundry.org) recommends the use of the Basic Formal Ontology (BFO) (www.ifomis.org/bfo) as a top-level ontology for biomedical ontologies. The ontology of Drug Discovery Investigations (DDI) has been developed to support automated drug discovery investigations run by a Robot Scientist "Eve" (Qi et al. 2010). DDI defines the essential entities for the recording and reasoning with data about the biological activity of compounds. A logically consistent description of the data and knowledge in the project is essential for the automation of scientific discovery (King et al. 2009).

DDI has been designed to be easily extendible to and compatible with other applications. Consequently DDI uses BFO and the RO (Relations Ontology) as design templates, and extends BFO by classes for drug design. For example the class *role* has been extended by the subclasses *drug role*, *agonist role*, etc. Some imported BFO classes were left unused, e.g., the class *connected_temporal_region*. "Unused" means that there are no child classes introduced, and no object properties added to an imported class.

Since ontologies intend to facilitate the representation of semantics for humans and computers, it is important that the navigation through an ontology by humans and during automated reasoning be efficient. Therefore unused classes diminish the usability of the ontology, because they unnecessarily complicate it. The simplification of the DDI by the removal of even a few classes has a considerable impact on efficiency. Eve runs thousands of parallel experiments and records millions of data items. Therefore any unused classes should be hidden from the data recording procedures as soon as the domain ontology has reached a stable state. At the current state-of-the-art this is not standard practice, because there is no easy way to identify and hide such classes. Reasoners do not report them as problematic. Ontology evaluation criteria are not explicit in regard to unused classes.

To be reliably usable, ontologies need to go through a Quality Assurance (QA) process, e.g. as a part of a larger software system. QA may involve an auditing regimen for discovering modeling errors and inconsistencies in the ontology. Without such an auditing process, the errors in an ontology may cause a malfunction of an information system using the ontology. A part of QA should be the hiding of imported classes that are not used by the domain ontology. We refer to this process as "hiding the redundant imports."

One of the methodologies proven to support QA is based on the automatic derivation of an Abstraction Network (AN) from an original ontology (Wang et al. 2007). An AN of an ontology is a compact secondary network summarizing the structure and content of the ontology. ANs were shown to support the identification of sets of concepts with higher concentrations of errors than randomly chosen control sets (Halper et al. 2007). Focusing QA efforts on such sets increases the yield of QA personnel, measured in number of problems found and corrected per unit of time.

In this paper, an AN is derived for the DDI, based on domain-defined and restriction-defined object properties,

* To whom correspondence should be addressed: zh5@njit.edu

referred to as a *partial area taxonomy* (Ochs et al. 2013). The top node of this taxonomy is the *Entity* node, representing 81 classes (12.5% of all DDI classes) that do not have any object properties. Large groups of classes represented by a single node in an AN, especially by the top node, tend to have a high concentration of modeling errors (Ochs et al. 2012; Ochs et al. 2013). Upon reviewing the classes summarized by the top node, we discovered that they contain most BFO classes.

BFO has no object properties and most classes imported from BFO were not used in the DDI and should be hidden from the DDI. An algorithm for hiding this "redundant portion" of BFO from DDI will be discussed. As many as 18 (56%) of the 32 BFO classes in the top node will be hidden from a subsequent release of DDI. All ontologies importing a top-level ontology should employ this AN-based approach to QA, for hiding redundant top level classes.

2 METHODS

In previous research, two kinds of ANs, called *area taxonomies* and *partial area taxonomies* have been developed to support QA for Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (Wang et al. 2007). We have also derived taxonomies for OWL-based ontologies that use similar but not identical definitional elements for orientation and QA. The AN derivation and QA methodologies were successfully applied to the Ontology of Clinical Research (OCRe) (Ochs et al. 2012) and the Sleep Domain Ontology (SDO) (Ochs et al. 2013). An *area* is defined as the set of all classes that are explicitly defined or inferred as being exactly in the domains of a given set of object properties. The list of names of the object properties is used to name the area. In general, the object properties that define an area can all be "domain-defined," or all be "restriction-defined" or there may be a mix of both.

Areas are connected by child-of links derived from the underlying ontology's subclass links. A *root* of an area is defined as a class that has no parents in the same area (i.e. none of its parents share its set of object properties). An area may have more than one root. Every root of an area defines a *partial area*: a set of classes that includes this root and all its descendants in the same area. Just as areas, partial areas are connected by child-of links derived from underlying subclass links. Partial area A is a child-of partial area B if a parent (superclass) of A's root class resides in B.

Fig. 1 (a) provides an excerpt of 13 classes taken from the DDI, along with five object properties. Two object properties *is_conretized_as* and *is about* have explicit domains (in red), while three object properties *has_participant*, *has_specified_input*, and *has_specified_output* are used in class restrictions (in black). Classes that are within the domain of a particular set of object properties are shown in a dashed bubble, e.g., the class *generally_dependent_continuant* is in the domain of the object property *is_concretized_as*. *Information content entity* is explicitly defined as the domain of *is about*, but it also inherits *is_concretized_as* from *generally_dependent_continuant*. *Conformation* and *contact information* are both implicitly in the domain of *is about* and *is_concretized_as* due to inheritance from *Information content entity*.

Fig. 1(b) shows the area taxonomy for the excerpt of DDI in Fig. 1(a). *Generally_dependent_continuant*, within the domain of *is_concretized_as*, is represented by the area with the name "is_concretized_as." Child-of links are shown as lines connecting the areas. Areas are organized into color-coded levels based-on their numbers of object properties. Areas with more object properties are lower down in the diagram.

Fig. 1(c) shows the partial area taxonomy for Fig. 1(a). Partial areas are represented by white boxes within area boxes. Each partial area is named by its root. The number of classes (including the root) in a partial area is shown in parentheses. For example, in the area named "has_specified_input, has_specified_output", there are two partial areas "data item extraction from journal article" and "documenting," each containing one class. We derived the partial area taxonomy for the DDI *Entity* hierarchy (Fig. 2).

Our methodology hides all the classes that were imported from a top-level ontology T into a domain ontology

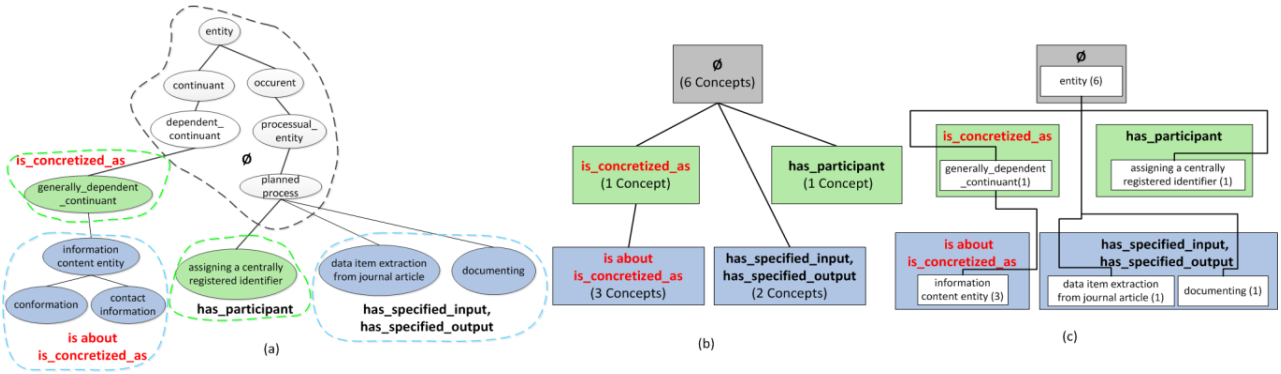

Fig. 1. (a) An excerpt of 13 classes and 5 object properties taken from the Ontology for Drug Discovery Investigations. **(b)** The area taxonomy derived from the classes in (a). **(c)** The partial area taxonomy derived from the classes in (a).

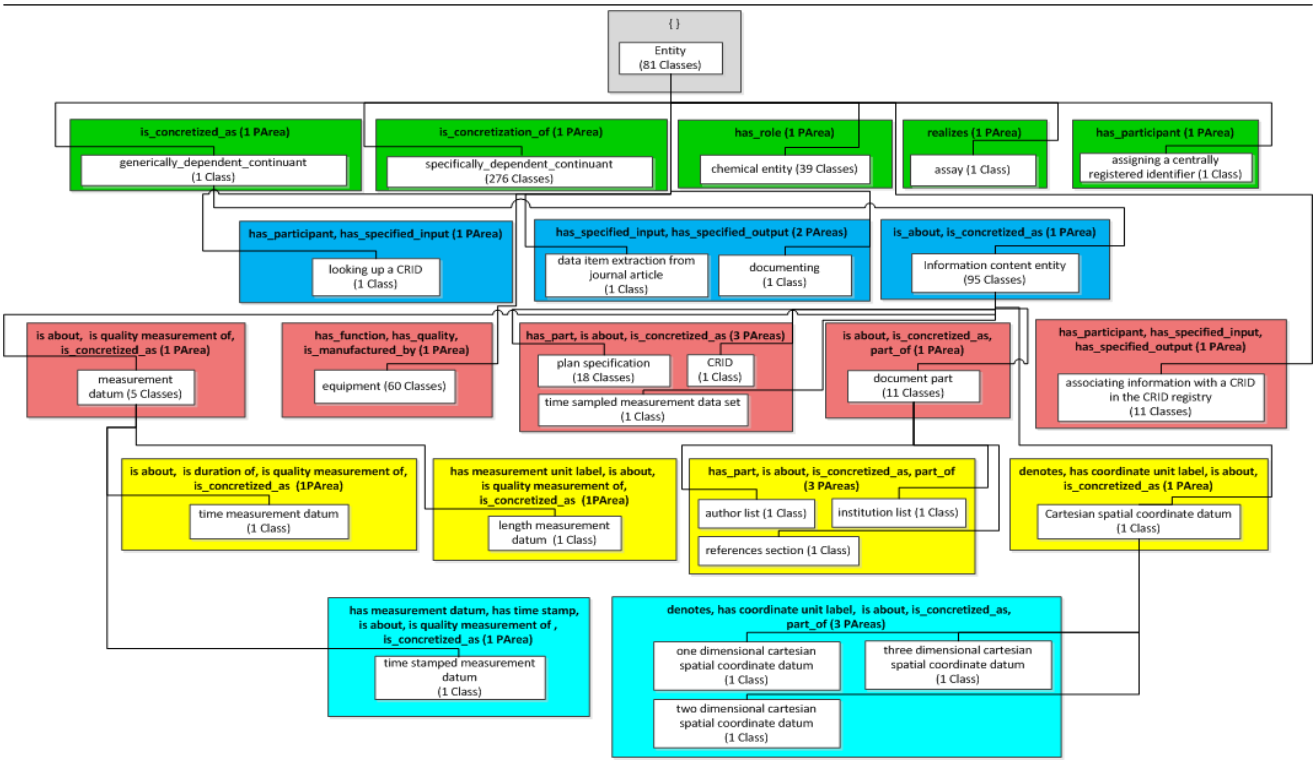

Fig. 2. Domain-defined and restriction-defined partial area taxonomy for the DDI's *entity* hierarchy

O, but were not utilized in O. In this paper, we are only considering top-level ontologies without object properties. Hence, a class C of T can be reused when constructing O by either associating object properties in O with C, or by modeling new classes of O as children or descendants of C. By limiting ourselves to classes of O in the root partial area R in the taxonomy of O, we are automatically limiting our attention to classes imported from T without any object properties in O. If a class of T has an object property in O (there are a seven such classes of BFO in DDI) then such a class cannot exist in the root partial area R of the taxonomy.

In addition, if a class in R was imported from T and is a leaf in O then, it was not utilized as the parent of any domain-related classes in O. Hence, such classes of O (T-imported classes, for short) are irrelevant for O and should be hidden. Following BFO, DDI and many BioPortal ontologies, we furthermore assume, in this paper, that no classes of T and O have multiple parents. In other words, the hierarchies of T and O are trees, which simplifies algorithmic processing.

We describe (but do *not* show) a recursive algorithm *Hide*(R, O, T, v) to automate the process of hiding unused T-imported classes from O. The algorithm uses three different contexts O, R and T, and operates in these contexts. *Hide* performs a post-order traversal of the classes in R. First, *Hide* is recursively applied to all subclasses w of the argument class v in R, initiating at R's root. When the traversal backtracks to v, the algorithm checks whether v is a leaf in O. Note that v may have been a leaf in O before, or may have become a leaf due to the removal of all its subclasses in O. Furthermore, if v was T-imported into O and is not in the range of an object property of O, then it is not used at all in O and needs to be hidden from O. By limiting the traversal to R, *Hide* is efficient with O(|R|) complexity, where R is only a small part of O, since the post-order traversal of a tree hierarchy is linear in the number of its nodes. However, the test whether a class is internal is done in O and not in R, because a leaf in R with subclasses in O was reused in O's modeling and should not be hidden from O.

3 RESULTS

DDI's *Entity* hierarchy contains 614 classes, which is 97.8% of DDI. There are 24 object properties, 14 of which are used within restrictions, 13 are given explicitly defined domains, and three have both.

First, we utilized the partial area taxonomy derivation methodology described by Ochs (Ochs et al. 2013) to derive the domain-defined or restriction-defined taxonomy of DDI's *Entity* hierarchy (614 classes) (Fig. 2). It contains 27 partial areas in 20 areas including five large partial areas (over 20 classes) and four medium size partial areas (of 5-20 classes). The other 18 partial areas include one class each. By reviewing the nine large and medium partial areas, e.g. *chemical entity* (39), *information content entity* (95), and *document part* (11), the user of DDI obtains a summary of the nature of classes in DDI. The names and sizes of these nine partial areas communicate knowledge about their content, supporting user orientation into DDI. The taxonomy

also displays the interaction among 18 partial areas of one class each, for sophisticated users, e.g. DDI curator, who are interested in orientation into the fine details of DDI's content and structure.

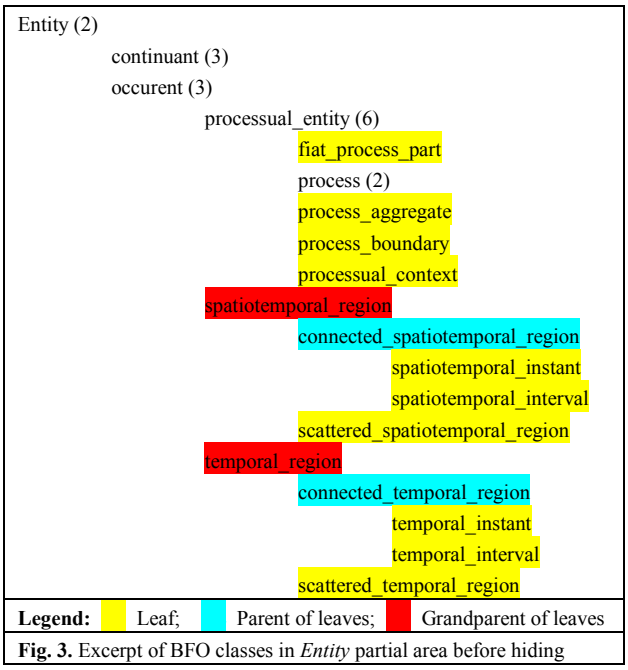

Fig. 3. Excerpt of BFO classes in *Entity* partial area before hiding

Finally, we executed the algorithm *Hide*(*Entity*, DDI, BFO, *entity*), where *entity* is the root class of the root partial area *Entity*, which is the partial area the algorithm traverses.

Out of 81 classes of the partial area *Entity*, 32 are BFO-imported. An excerpt of these BFO classes, focusing on descendants of *occurent,* is shown in an indented format in Fig. 3**.** Seven other BFO-imported classes have object properties added in DDI and thus should not be hidden. Four external subclasses of *spatial region* appear in ranges of object properties, and should also not be hidden. A number in parentheses in Fig. 3 indicates the number of children in O.

The 10 BFO-imported classes highlighted in yellow do not have any added DDI class descendants and are not used in range specifications of object properties. After these classes are hidden, the two classes in blue will appear as leaves and are also hidden. Then, the classes in red are hidden when they will appear as leaves. The curator of DDI (co-author LS) will hide the 18 unused BFO classes in a new release of DDI once the hiding mechanism is supported by BioPortal. There are only 14 reused classes from BFO in the *Entity* partial area that remain after hiding the 18 unused classes**.**

4 DISCUSSION

Reuse of a top-level ontology is quite common in BioPortal, i.e., at least 36 ontologies contain BFO classes. Since an ontology designer often does not know which top-level ontology classes will later be used, the common practice is to import the whole ontology. However, once a domain ontology is mature, there is no efficient way to remove unused classes. Our methodology hides the unused imports assuming that both ontologies have a tree hierarchy. We use a partial area taxonomy to limit the input size of the algorithm.

Following (www.imbi.uni-freiburg.de/ontology/), we distinguish between top-level, top-domain and domain ontologies. Reuse is practiced at two levels. Some top-domain ontologies such as BioTop (Beisswanger et al. 2008) and OGMS (bioportal.bioontology.org/ontologies/1414) are reusing BFO, while SDO (Arabandi 2010) reuses OGMS and thus indirectly also reuses BFO. There is a need to avoid a proliferation of unused imported classes.

In this paper, both the top level and the domain ontology have a tree hierarchy, but this is not always the case. OGMS has a tree hierarchy but SDO and the BioTop top-domain ontology do not. Hence, future research needs to consider cases where both ontologies do not have tree hierarchies. Furthermore, some top-domain ontologies such as BioTop, have object properties, another research consideration.

5 CONCLUSIONS

We described a recursive linear algorithm for hiding unused imported top-level ontology classes of an OWL-based ontology. The algorithm was demonstrated, hiding BFO-imported classes from the DDI.

ACKNOWLEDGMENTS

This work has been partially supported by the BRIEF award, Brunel University, London. We thank Natasha Noy of NCBO for her help.

REFERENCES

Arabandi, S. (2010). Developing a Sleep Domain Ontology. AMIA TBI/CRI Summt. San Francisco, CA.

Beisswanger, E., S. Schulz, et al. (2008). "BioTop: An Upper Domain Ontology for the Life Sciences." Appl Ontology **3**(4): 205-212.

Halper, M., Y. Wang, et al. (2007). "Analysis of error concentrations in SNOMED." AMIA Annu Symp Proc: 314-318.

King, R. D., J. Rowland, et al. (2009). "The automation of science." Science **324**(5923): 85-89.

Ochs, C., A. Agrawal, et al. (2012). "Deriving an abstraction network to support quality assurance in OCRe." AMIA Annu Symp Proc: 681-689.

Ochs, C., Z. He, et al. (2013). "Choosing the Granularity of Abstraction Networks for Orientation and Quality Assurance of the Sleep Domain Ontology". The 4th International Conference on Biomedical Ontology. Montreal, QC, Canada.

Qi, D., R. D. King, et al. (2010). "An ontology for description of drug discovery investigations." J Integr Bioinform **7**(3).

Wang, Y., M. Halper, et al. (2007). "Structural methodologies for auditing SNOMED." J Biomed Inform **40**(5): 561-581.

Aligning Pharmacologic Classes Between MeSH and ATC

Rainer Winnenburg[1], Laritza Rodriguez[1], Fiona Callaghan[1], Alfred Sorbello[2], Ana Szarfman[2], and Olivier Bodenreider[1]

[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA
[2]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, USA

ABSTRACT

Objective: To align pharmacologic classes in ATC and MeSH with lexical and instance-based techniques.

Methods: Lexical alignment: we map the names of ATC classes to MeSH through the UMLS, leveraging normalization and additional synonymy. Instance-based alignment: we associate ATC and MeSH classes through the drugs they share, using the Jaccard coefficient to measure class-class similarity. We use a metric to distinguish between equivalence and inclusion mappings.

Results: We found 221 lexical mappings, as well as 343 instance-based mappings, with a limited overlap (61). From the 343 instance-based mappings we classify 113 as equivalence mappings and 230 as inclusion mappings. A limited failure analysis is presented.

Conclusion: Our instance-based approach to aligning pharmacologic classes has the prospect of effectively supporting the creation of a mapping of pharmacologic classes between ATC and MeSH. This exploratory investigation needs to be evaluated in order to adapt the thresholds for similarity.

1 INTRODUCTION

The National Library of Medicine (NLM) and the Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) are collaborating on a research project to extract adverse drug reactions from the biomedical literature. More specifically, this investigation leverages the indexing of MEDLINE citations to extract associations between co-occurring drug entities and clinical manifestations in the context of adverse events.

The biomedical literature is indexed with the Medical Subject Headings (MeSH) vocabulary. For data mining purposes, however, adverse drug reactions are usually analyzed in reference to other standard vocabularies, namely the Anatomical Therapeutic Chemical (ATC) drug classification system for drug entities, and the Medical Dictionary for Regulatory Activities (MedDRA) for clinical manifestations. Toward this end, drug entities have to be mapped from MeSH to ATC, and manifestations from MeSH to MedDRA. This paper focuses only on the drug entities.

Drug entities include not only individual drugs (e.g., *atorvastatin*), but also drug classes (e.g., *statins*). In previous work, we have mapped individual drugs between RxNorm (which includes MeSH drugs) and ATC (Bodenreider and Taft, 2013; Winnenburg and Bodenreider, 2012). In contrast, no mapping is available between pharmacologic classes in MeSH and in ATC. Moreover, unlike individual drugs, whose names are relatively standardized across vocabularies, pharmacologic classes exhibit greater variability, not only in their names, but also in granularity. For example, the drug *lisinopril* is classified as *Angiotensin-Converting Enzyme Inhibitors* in MeSH, but as *ACE inhibitors, plain* in ATC.

The objective of this study is to investigate various ontology matching techniques for aligning pharmacologic classes between MeSH and ATC. Such methods are expected to facilitate the curation of a mapping by experts. To our knowledge, this work represents the first effort to map pharmacologic classes between MeSH and ATC using a sophisticated instance-based alignment technique.

2 BACKGROUND

The general framework of this study is that of ontology alignment (or ontology matching). Various techniques have been proposed for aligning concepts across ontologies, including lexical techniques (based on the similarity of concept names), structural techniques (based on the similarity of hierarchical relations), semantic techniques (based on semantic similarity between concepts), and instance-based techniques (based on the similarity of the set of instances of two concepts). An overview of ontology alignment is provided in (Euzenat and Shvaiko, 2007).

The main contribution of this paper is not to propose a novel technique, but rather to apply existing techniques to a novel objective, namely aligning pharmacologic classes between MeSH and ATC. To this end, we use lexical and instance-based techniques, because the names of pharmacologic classes and the list of drugs that are members of these classes are the main two features available in these resources.

2.1 Lexical techniques

Lexical techniques for ontology matching compare concept names across ontologies. When synonyms are available, they can be used to identify additional matches. Matching techniques beyond exact match utilize edit distance or normalization to account for minor differences between concept names.

As part of the Unified Medical Language System (UMLS), linguistically-motivated normalization techniques have been developed specifically for biomedical terms (McCray, et al., 1994). UMLS normalization abstracts away from inessential

* To whom correspondence should be addressed: obodenreider@mail.nih.gov

differences, such as inflection, case and hyphen variation, as well as word order variation. The UMLS normalization techniques form the basis for integrating terms into the UMLS Metathesaurus, but can be applied to terms that are not in the UMLS. For example, the ATC class *Thiouracils* (*H03BA*) and the MeSH class *Thiouracil* (*D013889*) match after normalization (ignoring singular/plural differences).

Lexical techniques typically compare the names of concepts across two ontologies as provided by these ontologies. However, additional synonyms can be used, for example, synonyms from the UMLS Metathesaurus. In other words, we leverage cosynonymy similarity for matching pharmacologic classes. In this case, although the ATC class *Anticholinesterases* (*N06DA*) and the MeSH class *Cholinesterase Inhibitors* (*D002800*) do not match lexically, both names are cosynonyms, because they are found among the synonyms of the UMLS Metathesaurus concept *C0008425*.

2.2 Instance-based techniques

Also called extensional techniques, instance-based techniques compare classes based on the sets of individuals (i.e., instances) of each class. Many biomedical ontologies consist of class hierarchies, but do not contain information about instances. Here, however, individual drugs (e.g., *atorvastatin*) are the members – not subclasses – of pharmacologic classes (e.g., *statins*). In other words, pharmacologic classes have individual drugs as instances, not subclasses.

Several methods have been proposed to implement instance-based matching. (Isaac, et al., 2007) decompose these methods into three basic elements: (1) A measure is used for evaluating the association between two classes based on the proportion of shared instances. Typical measures include information-based measures (e.g., Jaccard similarity coefficient) and statistical measures (e.g., log likelihood ratio). (2) A threshold is applied to the measures and pairs of classes for which the measure is above the threshold are deemed closely associated and mapping candidates. (3) Hierarchical relations in the two ontologies to be aligned can also be leveraged by deriving instance-class relations between instances of a given class and the ancestors of this class. In other words, in addition to asserted classes (i.e., the classes of which individual drugs are direct members), we also consider inferred classes (i.e., the classes of which asserted classes are subclasses). For example, the class asserted in MeSH for the drug *atorvastatin* is *Hydroxymethylglutaryl-CoA Reductase Inhibitors* (i.e., *statins*), whose parent concepts include *Anticholesteremic Agents*. Therefore, the class *Anticholesteremic Agents* is an inferred pharmacologic class for *atorvastatin*.

2.3 Related work

As part of the EU-ADR project, (Avillach, et al., 2013) extracted adverse drug reactions from the biomedical literature and mapped MeSH drugs to ATC through the UMLS. However, their mapping was limited to individual drugs and did not include pharmacologic classes.

Lexical techniques are a component of most ontology alignment systems (Euzenat and Shvaiko, 2007). While there have been attempts to map individual drugs from ATC to concepts in the UMLS and MeSH through lexical techniques, (Merabti, et al., 2011) note that these techniques are not appropriate for the mapping of pharmacologic classes.

While ***instance-based techniques*** are also available in many systems, the applicability of this technique is limited, because there is often no available information about instances as part of the ontologies to be aligned. For example, most biomedical terminologies and ontologies are simple class hierarchies. The instances of these classes are present in electronic medical record systems and clinical data warehouses, but typically not distributed along with the ontologies. One exception in the biomedical domain is the Gene Ontology (GO) (Ashburner, et al., 2000), for which the gene products annotated to GO terms can be considered instances of the corresponding classes. (Kirsten, et al., 2007) have aligned GO terms across the three hierarchies of GO through the gene products to which they are co-annotated.

To our knowledge, our work is the first attempt to align pharmacologic classes with instance-based techniques (i.e., beyond name matching), and the first application of aligning pharmacologic classes in ATC and MeSH. Moreover, while most ontology alignment systems mainly consider matches between equivalent classes, we are also interested in identifying those cases where one class is included in another class.

3 MATERIALS

3.1 Anatomical Therapeutic Chemical Drug Classification System (ATC)

The ATC is a clinical drug classification system developed and maintained by the World Health Organization (WHO) as a tool for drug utilization research to improve quality of drug use (ATC, 2013). The system is organized as a hierarchy that classifies clinical drug entities at five different levels: 1st level anatomical (e.g., *A*: *Alimentary tract and metabolism*), 2nd level therapeutic (e.g., *A10*: *Drugs used in diabetes*), 3rd level pharmacological (e.g., *A10B*: *Blood glucose lowering drugs, excluding insulins*), 4th level chemical (e.g., *A10BA*: *Biguanides*), and 5th level chemical substance or ingredient (e.g., *A10BA02*: *metformin*). The 2013 version of ATC integrates 4,516 5th-level drugs and 1,255 drug groups (levels 1-4).

3.2 MeSH

The Medical Subject Headings (MeSH) is a controlled vocabulary produced and maintained by the NLM (NLM, 2013). It is used for indexing, cataloging, and searching the

biomedical literature in the MEDLINE/PubMed database, and other documents. The MeSH thesaurus includes 26,853 descriptors (or "main headings") organized in 16 hierarchies (e.g., *Chemical and Drugs*). Additionally, MeSH provides about 210,000 supplementary concept records (SCRs), of which many represent chemicals and drugs. Each SCR is linked to at least one descriptor. While most chemical descriptors provide a structural perspective on drugs, some descriptors play a special role as they can be used to denote pharmacological actions in drug descriptors and SCRs. MeSH 2013 is used in this study.

3.3 RxNorm

RxNorm is a standardized nomenclature for medications produced and maintained by the U.S. National Library of Medicine (NLM) (NLM, 2013). RxNorm concepts are linked by NLM to multiple drug identifiers for commercially available drug databases and standard terminologies, including MeSH. RxNorm serves as a reference terminology for drugs in the US. The March 2013 version of RxNorm used in this study integrates about 10,500 base and salt ingredients. NLM also provides an application programming interface (API) for accessing RxNorm data programmatically (NLM, 2013).

3.4 Unified Medical Language System (UMLS)

The UMLS is a terminology integration system created and maintained by the National Library of Medicine (NLM) (NLM, 2013). The UMLS Metathesaurus integrates over 150 terminologies, including MeSH, but not ATC. Synonymous terms across terminologies are grouped into concepts and assigned the same concept unique identifier. The Metathesaurus provides a comprehensive set of synonyms for biomedical concepts and is often used for integrating terminologies beyond its own. NLM provides an application programming interface (API) for accessing UMLS data programmatically. Version 2012AB of the UMLS is used in this study.

4 METHODS

Our approach to aligning pharmacologic classes between MeSH and ATC based on their instances is depicted in Figure 1 and can be summarized as follows. First, we established a lexical alignment of MeSH and ATC classes based on the class names and their synonyms (Figure 1, right). We then constructed an instance-based alignment of MeSH and ATC classes considering the individual drugs shared by the classes (Figure 1, left). We mapped individual drugs from MeSH and ATC via their ingredients (IN) or precise ingredients (PIN) in RxNorm. We used a similarity measure and thresholds to identify class mappings and compared them with the mappings retrieved by the lexical approach.

In our alignment work, we excluded the 14 ATC groups of level 1 (anatomical classification), because they are too broad classes. We also excluded 164 of the 1,241 ATC groups (2^{nd} – 4^{th} level) corresponding to drug combinations, because combination drugs are often underspecified in ATC.

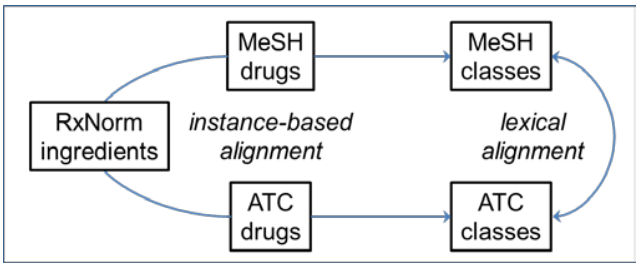

Figure 1. Alignment of ATC and MeSH classes, alignment via their instances (left) in comparison to direct lexical mapping of the class names (right).

Similarly, in MeSH, we excluded the top-level descriptors of the Chemicals and Drugs hierarchy (i.e., D01 - D27), as well as the top-level of the pharmacological action descriptors (*Pharmacologic Actions*, *Molecular Mechanisms of Pharmacological Action*, *Physiological Effects of Drugs*, and *Therapeutic Uses*).

4.1 Lexical alignment

We mapped all 1,077 eligible ATC classes (2^{nd} – 4^{th} level) to MeSH descriptors in the Chemicals and Drugs [D] tree using the UMLS Terminology Services (UTS). More precisely, we used the *ExactString* and *NormalizedString* search function of the UTS API 2.0 to establish mappings from the names of the ATC classes to UMLS concepts. We used normalization only when the exact technique did not result in a mapping. We then mapped the UMLS concepts to MeSH descriptor IDs.

4.2 Instance-based alignment

Mapping ATC drugs to RxNorm ingredients. In previous work we have mapped ATC single-ingredient drugs to Ingredients (IN) and Precise Ingredients (PIN) in RxNorm using a lexical approach with additional normalization steps (Winnenburg and Bodenreider, 2012). We used these mappings to establish the alignment of ATC and RxNorm drugs in this study.

Mapping MeSH drugs to RxNorm ingredients. Since MeSH drugs are integrated in RxNorm, mappings to equivalent drug concepts from MeSH can be obtained via the *getProprietaryInformation* function from the RxNorm API. We systematically exploited this information for all Ingredients (IN) and Precise Ingredients (PIN) in RxNorm and created a mapping table between RxNorm CUIs and MeSH Main Headings (MH) and Supplementary Concept Records (SCR).

Inferring class membership in ATC. We considered the hierarchical relations from 5^{th} level drugs to their 4^{th} level

chemical groups as asserted drug class membership. We inferred membership between 5th level drugs and groups of level 3 and 2 through transitive closure. For example, *temafloxacin* (J01MA05) is a member of the chemical group *Fluoroquinolones* (J01MA - asserted), the pharmacological group *QUINOLONE ANTIBACTERIALS* (J01M - inferred), and the therapeutic group *ANTIBACTERIALS FOR SYSTEMIC USE* (J01 - inferred).

Table 1. Asserted and inferred MeSH classes for the drug *temafloxacin* (*C054745*) with type of relationship to the drug and tree numbers in MeSH.

Type	Asserted Classes	Inferred Classes
PA	***Anti-Bacterial Agents*** (D000900) [D27.505.954.122.085]	***Anti-Infective Agents*** (D000890) [D27.505.954.122]
MH	***Fluoroquinolones*** (D024841) D03.438.810.835.322	***Quinolones*** (D015363) [D03.438.810.835]
		Quinolines (D011804) [D03.438.810]
		Heterocyclic Compounds, 2-Ring (D006574) [D03.438]

Inferring class membership in MeSH. We associated each RxNorm ingredient (IN or PIN) with its corresponding MeSH supplementary concept record (SCR) or main heading (MH). In turn, we associated these drugs with their asserted classes. For an SCR, we considered its pharmacological actions, as well as the MeSH heading(s) mapped to. For a MH, we considered its pharmacological actions, as well as its direct ancestors. These constitute the asserted classes. We inferred membership between the drugs and higher-level descriptors in the MeSH hierarchy. For example, as shown in Table 1, the SCR *temafloxacin* has *Anti-Bacterial Agents* as pharmacological action and *Fluoroquinolones* as main heading mapped to. Form these asserted classes, we infer membership to *Anti-Infective Agents* (from *Anti-Bacterial Agents*) and to *Quinolones*, *Quinolines*, and *Heterocyclic Compounds, 2-Ring* (from *Fluoroquinolones*).

Measure for aligning ATC and MeSH classes. Based on the asserted and inferred class membership of drugs in ATC and MeSH we conducted a pairwise comparison of all ATC against all MeSH classes. For each pair of ATC class (A) and MeSH class (M), we computed the Jaccard coefficient. In order to reduce the similarity of pairs of classes with a small number of shared members, we used a modified version of the Jaccard coefficient, JCmod, as suggested in (Isaac, et al., 2007),

$$\mathrm{JC}(A,M) = \frac{|A \cap M|}{|A \cup M|}$$

$$\mathrm{JCmod}(A,M) = \frac{\sqrt{|A \cap M| \times (|A \cap M| - 0.8)}}{|A \cup M|}$$

where $A \cap M$ represents the number of drugs common to A and M, and $A \cup M$ the total number of unique drugs in both classes.

The Jaccard coefficient measures the similarity between the two classes, but does not reflect whether one class is included in the other. Because of the difference in granularity between classes in ATC and MeSH, we introduce a simple metric for detecting whether the drugs that are not shared by both classes are primarily in one of the two classes. This "one-sidedness" coefficient is calculated as follows:

$$0, \quad \text{for } a = 0 \text{ and } m = 0$$

$$|a-m| / a+m, \quad \text{otherwise.}$$

where *a* and *m* are the number of drugs specific to the ATC class and the MeSH class, respectively. Thus, a "one-sidedness" coefficient close to 0 indicates that the drugs that are not shared by the two classes are evenly distributed between the ATC and MeSH class. In contrast, a coefficient close to 1 indicates that only one of the classes contains most of the drugs that are not shared by the other.

Thresholds. In order to select the best equivalent or inclusion mappings between ATC and MeSH, we characterize each pair of ATC and MeSH classes with respect to Jaccard similarity and one-sidedness. Low one-sidedness indicates equivalence and high one-sidedness indicates inclusion. High Jaccard similarity indicates strong overlap between the two classes. Based on preliminary analysis, we selected of a threshold of 0.5 for the one-sidedness metric. Similarly, we selected of a threshold of 0.5 and 0.25 for Jaccard similarity for equivalence (low one-sidedness) and inclusion (high one-sidedness), respectively. The lower threshold for Jaccard similarity for inclusion was determined empirically. As shown in Table 2, each pair of ATC and MeSH classes is characterized as an equivalence mapping (EQ+), an inclusion mapping (IN+), or not a mapping (EQ- and IN-).

5 RESULTS

5.1 Lexical alignment

For the 1,077 eligible ATC groups of level 2-4, we were able to retrieve 226 mappings to descriptors from the Chemicals and Drugs [D] tree in MeSH. We have 18 mappings for therapeutic classes (2nd level), 42 for pharmacological classes (3rd level), and 161 for chemical classes (4th level). We ignored mappings for the broad anatomical classes (1st level). Of the 221 mappings, 96 are to pharmacological ac-

tions (functional perspective) in MeSH, whereas 125 are to other descriptors at various levels of the MeSH hierarchy (structural perspective).

5.2 Instance-based alignment

Of the 1,077 eligible ATC groups, 874 (81%) could be associated with at least one descriptor or pharmacological action in MeSH. We identified a total of 933 associations for the 874 ATC groups (multiple associations per ATC group possible). As shown in Table 2, based on the one-sidedness metric, we characterized 323 associations as equivalence and 610 as inclusion. Of the 323 equivalence associations, 113 (35%) exhibit high Jaccard similarity and are selected as equivalence mappings (EQ+). Of the 610 inclusion associations, 230 (38%) exhibit high Jaccard similarity and are selected as inclusion mappings (IN+). The other associations (EQ- and IN-) are not deemed strong enough to denote mappings. In summary, we were able to characterize as a mapping (EQ+ and IN+) 343 (37%) of the associations between ATC and MeSH classes. It should be mentioned that we were not able to obtain mappings to MeSH classes for 203 ATC classes, because they only contain drug instances that could not be mapped to drugs in MeSH.

Table 2. Characterization of the associations between ATC and MeSH classes based on Jaccard similarity and score for one-sidedness. The numbers in grey fields indicate the associations that are not strong enough to denote mappings.

		One-sidedness		
		≥ .5	< .5	Total
Jaccard	≥ .5	IN+ (230)	EQ+ (113)	343
	[.25-.5[		EQ- (210)	590
	< .25	IN- (380)		
	Total	610	323	933

5.3 Comparison between lexical and instance-based alignment

As illustrated in Table 3, from the 221 lexical mappings between ATC and MeSH classes, we could confirm 61 with our instance-based approach (30 as equivalence mappings, 31 as inclusion mappings). For 19 of the lexical mappings we found an association with low Jaccard similarity (IN- / EQ -), and for 141 of the lexical mappings we did not find any association through the instance-based alignment (mainly due to the lack of any mapping for the drug instances in these classes). Finally, the instance-based approach produced 282 additional drug class mappings that were not detected by the lexical approach, whereas 633 (571 + 62) ATC classes could neither be mapped by the lexical nor the instance-based approach.

Table 3. Comparison between lexical and instance-based alignment.

		Instance-based			
		Yes	No	No assoc.	Total
Lexical	Yes	61	19	141	221
	No	282	571	62	915
	Total	343	590	203	1136

6 DISCUSSION

6.1 Examples and failure analysis

True positive for equivalent instance-based mappings. We identify an equivalence mapping between the 4th-level ATC group *Fluoroquinolones* (J01MA) and the MeSH descriptor *Fluoroquinolones* (D024841). The two classes share 14 drugs. The ATC group has one extra drug (*moxifloxacin*), and the MeSH descriptor has 2 (*flumequine* and *besifloxacin*). Jaccard similarity is high (0.82) and the one-sidedness score is low (0.33), because the 3 drugs that are not in common are not all on the same side. This mapping is also identified by the lexical technique (exact match).

True positive for inclusion instance-based mappings. We identify an inclusion mapping between the 4th-level ATC group *Fluoroquinolones* (S01AE) and the MeSH descriptor *Fluoroquinolones* (D024841). Although the two classes are seemingly identical, our mapping is identified as an inclusion, with 7 drugs in common, 1 drug specific to the ATC class and 9 drugs specific to the MeSH class. In fact, the ATC class is not the same general class for anti-infective agents as in the example above (J01MA), but rather the specific class of fluoroquinolones for ophthalmic use (S01AE). The fluoroquinolones used for eye disorders are a subset of all fluoroquinolones and the ATC class S01AE is appropriately characterized as being included in the MeSH class for fluoroquinolones. This example also illustrates a false positive for the lexical mapping, since it is generally assumed that lexical mappings are equivalence mappings.

False negative for equivalent instance-based mappings. Many ATC and MeSH classes share only one or very few drugs, making it difficult to assess equivalence or inclusion. For example, the 4th-level ATC group *Silver compounds* (D08AL) and the MeSH descriptor *Silver Compounds* (D018030) share only one drug (silver). The modified version of the Jaccard coefficient has a score of 0.45 in this case, which is below our threshold of 0.5 for equivalence.

During this failure analysis, we discovered that some MeSH drugs did not have a pharmacological action assigned to them as we expected. For example, while *pyrantel* is listed as *Antinematodal Agents*, *oxantel* is not. We are investigating whether the pharmacological action for this SCR should be inferred from the descriptor to which it is mapped (*Pyrantel* in this case). Because of these missing pharmacologic

actions, the 3rd-level ATC group *ANTINEMATODAL AGENTS* (P02C) fails to be mapped to the MeSH pharmacological action *Antinematodal Agents* (D000969), the Jaccard similarity being just below the threshold (0.49).

Discrepancy between lexical and instance-based alignment (missed lexical mapping). Despite the use of UMLS synonymy and normalization, the lexical alignment fails to identify a mapping between the 3rd-level ATC group *POTASSIUM-SPARING AGENTS* (C03D) and the MeSH pharmacological action *Diuretics, Potassium Sparing* (D062865). In contrast, the instance-based alignment identifies an equivalence mapping with very high Jaccard similarity (0.99). This finding is consistent with the conclusions of (Merabti, et al., 2011).

Discrepancy between lexical and instance-based alignment (missed instance-based mapping). We have identified several causes for discrepancies between lexical and instance-based alignments. As mentioned earlier, some ATC classes only contain drugs that cannot be mapped to MeSH through RxNorm, which we used to bridge between the two. Sometimes, the best instance-based mapping is to another class than the class found by the lexical technique. Finally, some drugs entities and biologicals (e.g., vaccines) are less well standardized than common drugs. For this reason, the instance-based alignment is unable to map these classes, when simple lexical techniques can.

6.2 Limitations and future work

This exploratory investigation has several limitations, which we plan to address in future work.

Evaluation. This exploratory investigation focuses primarily on the methodology and feasibility of the alignment, and does not include a formal evaluation. Since ATC and MeSH pharmacological actions are being integrated into RxNorm, we will use the alignment created by RxNorm experts as the gold standard to evaluate our methods.

Perspective. Our perspective in this investigation is ATC-centric, because we consider the best MeSH mapping for each ATC class, but not the best ATC mapping for each MeSH class. One future goal is to explore both directions using the same methodology.

Bias towards equivalence mappings. Because we restrict our exploration to the MeSH class with the best Jaccard similarity for each ATC class (which we subsequently categorize as equivalence or inclusion), and because of the differential threshold for Jaccard similarity between equivalence (0.5) and inclusion mappings (0.25), we potentially fail to consider a good inclusion mapping (e.g., with a similarity score of 0.39 [> 0.25]), when the best MeSH class is a bad equivalent mapping (e.g., with a similarity score of 0.41 [< 0.5]).

6.3 Significance

To our knowledge, our work is the first attempt to align pharmacologic classes with instance-based techniques, distinguishing between equivalence and inclusion relations, as well as the first application of alignment between pharmacologic classes in ATC and MeSH. Our instance-based approach to aligning pharmacologic classes has yielded 343 mappings, and has the prospect of effectively supporting the creation of a mapping of pharmacologic classes between ATC and MeSH. This exploratory investigation needs to be evaluated in order to adapt the thresholds for similarity.

ACKNOWLEDGEMENTS

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine and by the Center for Drug Evaluation and Research of the Food and Drug Administration. The authors want to thank Rave Harpaz and Anna Ripple for useful discussions.

DISCLAIMER

The findings and conclusions expressed in this report are those of the authors and do not necessarily represent the views of the FDA.

REFERENCES

Ashburner, M., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, **25**, 25-29.

Anatomical Therapeutic Chemical (ATC) classification: http://www.whocc.no/atc/

Avillach, P., *et al.* (2013) Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project, *J Am Med Inform Assoc*, **20**, 446-452.

Bodenreider, O. and Taft, L.M. (2013) A mapping of RxNorm to the ATC/DDD Index helps analyze US prescription lists, *AMIA Annu Symp Proc*, (submitted).

Euzenat, J. and Shvaiko, P. (2007) *Ontology matching*. Springer, New York.

Isaac, A., *et al.* (2007) An empirical study of instance-based ontology matching. In Aberer, K., *et al.* (eds), *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference (ISWC'07/ASWC'07)*. Springer-Verlag, pp. 253-266.

Kirsten, T., Thor, A. and Rahm, E. (2007) Instance-based matching of large life science ontologies In Cohen-Boulakia, S. and Tannen, V. (eds), *Data Integration in the Life Sciences: 4th International Workshop, DILS 2007, Philadelphia, PA, USA*. Springer, pp. 172-187.

McCray, A.T., Srinivasan, S. and Browne, A.C. (1994) Lexical methods for managing variation in biomedical terminologies, *Proc Annu Symp Comput Appl Med Care*, 235-239.

Merabti, T., *et al.* (2011) Mapping the ATC classification to the UMLS metathesaurus: some pragmatic applications, *Stud Health Technol Inform*, **166**, 206-213.

Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/

RxNorm: http://www.nlm.nih.gov/research/umls/rxnorm/

RxNorm API: http://rxnavdev.nlm.nih.gov/RxNormAPI.html

Unified Medical Language System (UMLS): https://uts.nlm.nih.gov/

Winnenburg, R. and Bodenreider, O. (2012) Mapping drug entities between the European and American standards, ATC and RxNorm, *Poster Proceedings of the Eighth International Conference on Data Integration in the Life Sciences (DILS 2012)*, 22.

Analysis of Vaccine-related Networks using Semantic MEDLINE and the Vaccine Ontology

Yuji Zhang[1,*], Cui Tao[1], Yongqun He[2], Pradip Kanjamala[1], Hongfang Liu[1]
[1] Department of Health Sciences Research, Mayo College of Medicine, Rochester, MN 55905, USA
[2]Unit of Laboratory of Animal Medicine, University of Michigan, Ann Arbor, MI 48109, USA

ABSTRACT

A major challenge in the vaccine research has been to identify important vaccine-related networks and logically explain the results. In this paper, we showed that network-based analysis of vaccine-related networks can discover the underlying structure information consistent with that captured by the Vaccine Ontology and propose new hypotheses for vaccine disease or gene associations. First, a vaccine-vaccine network was inferred using a bipartite network projection strategy on the vaccine-disease network extracted from the Semantic MEDLINE database. In total, 76 vaccines and 573 relationships were identified to construct the vaccine network. The shortest paths between all pairs of vaccines were calculated within the vaccine network. The correlation between the shortest paths of vaccine pairs and their semantic similarities in the Vaccine Ontology was then investigated. Second, a vaccine-gene network was also constructed, in which several important vaccine-related genes were identified. This study demonstrated that a combinatorial analysis using literature knowledgebase, semantic technology, and ontology is able to reveal unidentified important knowledge critical to biomedical research and public health and generate testable hypotheses for future experimental verification.

1 INTRODUCTION

Vaccines have been one of the most successful public health interventions to date with most vaccine-preventable diseases having declined in the United States by at least 95-99% (1994). However, vaccine development is getting more difficult as more complex organisms become vaccine targets. In recent years, drug repositioning has been growing in last few years and achieved a number of successes for existing drugs such as Viagra (Goldstein, Lue et al. 1998) and thalidomide (Singhal, Mehta et al. 1999). By definition, drug repositioning is the "process of finding new users outside the scope of the original medical indications for existing drugs or compounds" (Chong and Sullivan 2007). In 2009, more than 30% of the 51 new medicines and vaccines were developed based on previous marketed products. This suggested that drug repositioning has drawn great attention from the both industry and academic institutes (Graul, Sorbera et al. 2010). However, many of these drug repositioning have been serendipitous discoveries (Ashburn and Thor 2004) or on observable clinical phenotypes, which are lack of systematic ways to identify new targets. Recent research has shown that bioinformatics-based approaches can aid to reposition drugs based on the complex relationships among drugs, diseases and genes (Liu, Fang et al. 2013). Such approaches can also be applied in the future vaccine development.

In recent years, high-throughput biological data and computational systems biology approaches has provided an unprecedented opportunity to understand the disease etiology and its underlying cellular subsystems. Biological knowledge such as drug-disease networks, and biomedical ontologies have accelerated the development of network-based approaches to understanding disease etiology (Ideker and Sharan 2008; Barabasi, Gulbahce et al. 2011) and drug action (network pharmacology) (Berger and Iyengar 2009; Mathur and Dinakarpandian 2011). Such approaches could also be applied in the vaccine research, aiming to investigate the vaccine-related associations derived from public knowledgebase such as PUBMED literature abstracts. For example, a Vaccine Ontology (VO)-based literature mining research was reported last year to study all gene interactions associated with fever alone or both fever and vaccine (Hur, Ozgur et al. 2012). This study focused on the retrieval of gene-gene associations based on their direct interactions in the context of fever and vaccine. The centrality-based network approach (Ozgur, Vu et al. 2008) evaluated the level of importance for each gene in extracted gene interaction network. Novel gene interactions were identified to be essential in fever- or vaccine-related networks that could not be found before. A similar VO and centrality-based literature mining approach was also used to analyse vaccine-associated IFN-γ gene interaction network (Ozgur, Xiang et al. 2011). Ball et al. compiled a network consisting of 6,428 nodes (74 vaccines and 6,354 adverse events) and more than 1.4 million interlinkages, derived from

* To whom correspondence should be addressed: Zhang.Yuji@mayo.edu

Vaccine Adverse Event Reporting System (VAERS) (Ball and Botsis 2011). This network demonstrated a scale-free property, in which certain vaccines and adverse events act as "hubs". Such network analysis approaches complement current statistical techniques by offering a novel way to visualize and evaluate vaccine adverse event data. However, the relationships among different vaccines in the context of vaccine-vaccine and vaccine-gene networks have not been well studied yet. A systematic level investigation of such relationships will help us understand better how vaccines are related to each other and whether such information can complement the existing knowledge such as VO.

To analyse the possible common protective immunity or adverse event mechanisms among different vaccines, it is critical to study all possible vaccine-vaccine and vaccine-gene associations using network analysis approaches. The hypotheses behind this are: (1) if two vaccines have coupling relationship with common disease(s)/gene(s), they are linked in the vaccine network; (2) the closer two vaccines are in the vaccine network, the more similar they are in the context of literature knowledgebase, such as Semantic MEDLINE (Rindflesch, Kilicoglu et al. 2011). In this paper, we proposed a network-based approach to investigate the underlying relationships among vaccines in the context of the vaccine-related network derived from Semantic MEDLINE. The distances of the vaccines were further compared with their semantic similarities in the VO. The results demonstrated that the structure information of vaccine network is consistent with that captured by VO. Such network-based analysis can serve as an independent data resource to construct and evaluate biomedical ontologies. In addition, the vaccine-gene network was also constructed based on Semantic MEDLINE information, in which important vaccine-related genes were identified and further investigated by VO and related independent resources.

The rest of the paper is organized as follows. Section 2 introduces the data resources and the proposed network-based framework. Section 3 illustrates the results generated from each step in the proposed computational framework. Section 4 provides a thorough discussion of the results and concludes the paper.

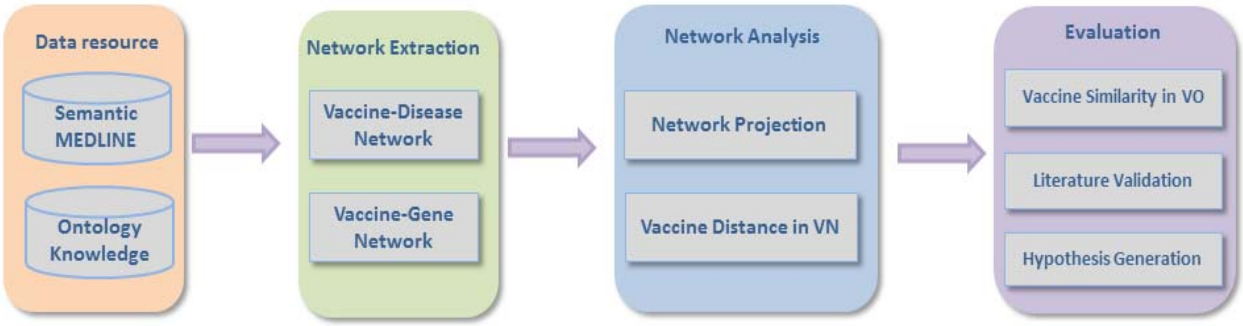

Fig. 1. Overview of the proposed framework. The proposed method consists of three steps: 1) Extraction of vaccine-related associations from Semantic MEDLINE using ontology based terms; 2) Network based analyses to identify vaccine-vaccine associations and vaccine-gene associations; 3) Evaluation of inferred vaccine-vaccine and vaccine-gene relationships using vaccine ontology hierarchical structure and literature validation.

2 MATERIALS AND METHODS

In this section, we describe the data resources and preprocessing method in this work. We then introduce our proposed network-based approach for investigating vaccine-related associations derived from PubMed literature abstracts. The evaluation of the discovered vaccine-vaccine and vaccine-gene relationships is conducted based on the VO hierarchy and logical definitions. Fig. 1 illustrates the steps of the proposed approach.

2.1 Data Resources and preprocessing

2.1.1. Data Resources

In this study, we use Semantic MEDLINE as the data resource to build the networks. Semantic MEDLINE (Rindflesch, Kilicoglu et al. 2011) is a National Library of Medicine (NLM) initiated project which provides a public available database that contains comprehensive resources with structured annotations for information extracted from more than 19 million MEDLINE abstracts. Since the Semantic MEDLINE is a comprehensive resource that contains heterogeneous data with different features extracted, our previous research has reorganized this data source and optimized it for informatics analysis (Tao, Zhang et al. 2012). Using the Unified Medical language System (UMLS) semantic types and groups (2012), we extracted unique associations among diseases, genes, and drugs, and represented them in six Resource Description Framework (RDF) graphs. In this paper, we used our optimized Semantic MEDLINE RDF data as the data source to perform network analysis for vaccine-related networks.

Our RDF-based Semantic MEDLINE resource currently contains 843k disease-disease, 111k disease-gene, 1277k disease-drug, 248k drug-gene, 1900k drug-drug, and 49k gene-gene associations. Since this resource contains high-level terms (e.g., gene, protein, disease) that are not useful for network analysis, we further manually filtered out these terms using the following strategy. For disease terms, we only included those terms that are included in ICD9. For gene terms, we only include those terms that have an Entrez gene ID.

2.1.2. Data extraction

We identified those associations relevant to vaccines only. Specifically, vaccine terms were identified based on SNOMED CT (http://www.ihtsdo.org/snomed-ct). All the terms under the SNOMED CT term *Vaccine* (CUI: C0042210) were first extracted. A manual review by 3 experts further removed those common terms (e.g., bacteria vaccine) or animal vaccine terms.

2.2 Network analysis of Vaccine Network

2.2.1. Projection of bipartite vaccine-disease network

In graph theory, a bipartite network is composed of two non-overlapping sets of nodes and links that connect one node in the first node set with one node in the second node set. The properties of bipartite networks are often investigated by considering the one-mode projection of the bipartite network. The one-mode projection network can be created by connecting two nodes in the same node set if they have at least one common neighboring node in the other node set. For instance, the vaccine-disease association network is one bipartite network: vaccines and diseases constitute two node sets, and links are generated between vaccine and disease if they are associated in the Semantic MEDLINE. Therefore, the vaccine-vaccine network can be investigated by projecting vaccine-disease associations to vaccine-vaccine associations, in which two vaccines are connected if they are associated with at least one same disease. In this work, all links were generated based on the associations extracted from Semantic MEDLINE as described in Section 2.1. A vaccine-vaccine network was generated consisting of all the links identified in vaccine-disease associations.

2.2.2. Network distance between vaccines

The distance between any two vaccines in the vaccine network was calculated as the length of the shortest path between them (Fekete, Vattay et al. 2009). The hierarchical clustering analysis was performed on the distance matrix of all vaccines (Guess and Wilson 2002). A heat map was generated based on the clustering analysis results.

2.2.3. Analysis of vaccine-gene network

The vaccine-gene network was constructed by vaccine-gene associations extracted from the drug-gene associations in our RDF-based data resource. The important vaccine-related genes were identified by their significant higher node degree compared to other vaccine/gene in the same network. The Cytoscape tool (Smoot, Ono et al. 2011) was selected to visualize the network. Cytoscape is an open-source platform for integration, visualization and analysis of biological networks. Its functionalities can be extended through Cytoscape plugins. Scientists from different research fields have contributed over 160 useful plugins so far. These comprehensive features allow us to perform thorough network level analyses and visualization of our association tables, and integration with other biological networks in the future.

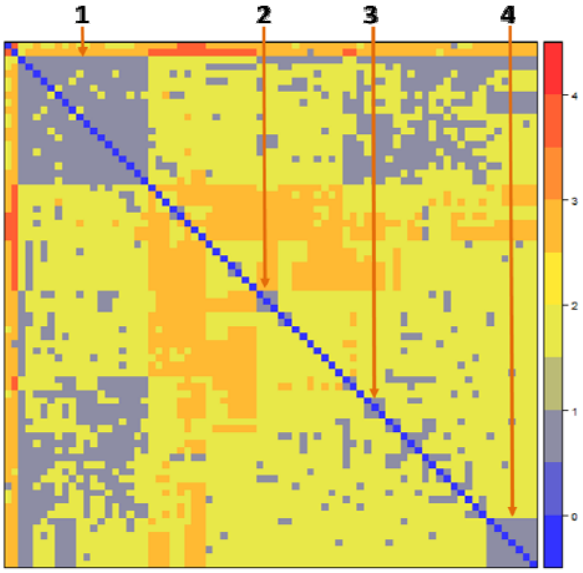

Fig. 2. The heat map of vaccine-vaccine associations. The shortest path matrix of all vaccine pairs was used to generate the heat map. Each row (column) represents a vaccine term. The color scale represents the shortest path between any vaccine pair.

2.3 Analysis of vaccine groups using VO

The community-based Vaccine Ontology (VO) has included over 4,000 vaccine-specific terms, including all licensed human and veterinary vaccines currently used in the USA. Logical axioms have been defined in VO to represent the relations among vaccine terms (Ozgur, Xiang et al. 2011). The Semantic MEDLINE analysis uses SNOMED terms to represent various vaccines. VO has established automatic mapping between SNOMED vaccine terms and VO terms. Based on the mapping, we first extracted all vaccine terms from the Semantic MEDLINE and mapped to VO. The ontology term retrieval tool OntoFox (Xiang, Courtot et al. 2010) was then applied to obtain the hierarchies of the total vaccines or subgroups of the vaccines identified in this study.

3 RESULTS

3.1 The overall network view

In total, 76 vaccines, annotated by the SNOMED CT term *Vaccine* (CUI: C0042210), were used to extract related vaccine-disease and vaccine-gene associations from the drug-disease and drug-gene association tables respectively. In the vaccine-disease network, there were 1127 nodes (178 vaccines and 949 diseases) and 1741 vaccine-disease associations. In the vaccine-gene network, there were 170 nodes (85 vaccines and 85 genes) and 94 vaccine-gene associations. One vaccine network was generated by the projection of the vaccine-disease bipartite network, consisting of 76 vaccines and 573 associations. This vaccine network was then used to analyze the vaccine relationships. The derived vaccine-gene network was also investigated by the VO knowledge.

3.2 Analysis of vaccine network

Fig. 2 showed a heat map of hierarchical analysis results, providing a direct visualization of potential vaccine-vaccine associations. Here we selected four relatively big vaccine-vaccine association groups on the diagonal from Fig. 2 and explain them in detail:

1) This group contains 18 very widely-studied vaccines. Many interesting results are obtained from the analysis of this group of vaccine-disease-vaccine associations. For example, the results from this group show that influenza vaccines and Rabies vaccines have been associated with the induction of a severe adverse event Guillain-Barré syndrome (GBS) (Hemachudha, Griffin et al. 1988; Hartung, Keller-Stanislawski et al. 2012). GBS is a rare disorder in which a person's own immune system damages their nerve cells, causing muscle weakness and sometimes paralysis. This group also includes five other vaccines associating with nervous system disorder, including Pertussis Vaccine (Wardlaw 1988), Diphtheria-Tetanus-Pertussis Vaccine (Corkins, Grose et al. 1991), Hepatitis B Vaccines (Comenge and Girard 2006), Chickenpox Vaccine (Bozzola, Tozzi et al. 2012), and Poliovirus Vaccine (Friedrich 1998; Korsun, Kojouharova et al. 2009). As shown by a VO hierarchical structure layout (Fig. 3), these seven vaccines belong to different bacterial or viral vaccines. The Diphtheria-Tetanus-Pertussis vaccine (DTP) is a combination vaccine that contains three individual vaccine components, including a Pertussis vaccine. DTP is asserted in VO as a subclass of "Diphtheria-Tetanus vaccine". Different from SNOMED, VO logically defines vaccines based on their relation to the pathogen organisms defined in the NCBI_Taxon ontology. Since multiple inheritances are not used in VO, an inference using an ontology reasoner was used to infer that the DTP is also *Bordetella pertussis* vaccine (*i.e.*, Pertussis vaccine) (Fig. 3). It is likely that the association of the combination vaccine DTP with neurological disorder is at least partially due to the Pertussis vaccine component.

Our study also identified many other diseases associating with different vaccines. For example, five vaccines (*e.g.*, pertussis vaccine) were found to be associated with various types of antimicrobial susceptibility, and eight vaccines (*e.g.*, influenza vaccine) have been co-studied with patients having the asthma condition. Due to the relative poor annotation of the vaccine data in the Semantic MEDLINE system, the vaccines identified in the semantic analysis were poorly classified. The incorporation of VO in the study clearly classifies these vaccines, leading to better understanding of the result of the Semantic MEDLINE analysis.

2) This group of vaccines, including Q fever vaccine, Parvovirus vaccine, and Tick-borne encephalitis vaccine, is associated with the common disease "Delayed Hypersensitivity". Delayed type reactions may occur at days after vaccination and often raise serious safety concerns. Delayed hypersensitivity is not antibody-mediated but rather is a type of cell-mediated response. The study of common vaccines and related gene and pathway features related to the delayed reaction will help to reveal the cause of DTR and eventually prevent it. While these vaccines are developed against different bacterial or viral diseases, there may be similarities among these vaccines, such as common vaccine ingredients (e.g., adjuvant) and a shared target to some common biological pathway in humans. An identification of these common features may indicate a common cause of the DTR.

3) This group of vaccines is associated with the common disease "Mumps". The vaccines in this group include Mumps Vaccine, "measles, mumps, rubella, varicella vaccine", and "diphtheria-tetanus-pertussis-haemophilus b conjugate vaccine" (DTP-Hib). The first two vaccines protect against Mumps. DTP-Hib was compared with a Mumps vaccine in a study (Henderson, Oates et al. 2004).

4) This vaccine group consists of seven vaccines (e.g., *Brucella abortus* vaccine and bovine rhinotracheitis vaccine) with direct associations between them. They are all associated with the common term "calve" in the literature abstracts. Since "calve disease" has a synonym "Scheuermann's Disease", these vaccines have all been linked to "Scheuermann's Disease". This is due to the ambiguity of the Nature Language Processing (NLP) process. This can be improved by future improvement of the disambiguity capacity of NLP tools.

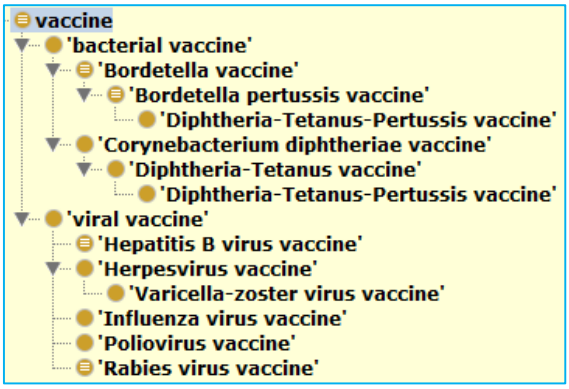

Fig. 3. The VO hierarchical structure of the seven vaccines associating with neurological disorder. A reasoning process assigned the Diphtheria-Tetanus-Pertussis vaccine under *Bordetella pertussis* vaccine. The Protégé-OWL editor 4.2 was used for the figure generation.

3.3 Vaccine-gene network

In the vaccine-gene network, many genes were found to interact with different vaccines. For example, our study identified that CD40LG (CD40 ligand) is closely associated with five vaccines: Diphtheria toxoid vaccine, Cholera vaccine, Tetanus vaccine, Chickenpox vaccine, and inactivated poliovirus vaccine (Fig. 4). CD40LG plays an important role in antigen presentation and stimulation of cytotoxic T lymphocytes (Kornbluth 2000). CD40LG can also be used in rational vaccine adjuvant design (Kornbluth and Stone 2006). Our finding confirms the important role of CD40LG and provides specific details on how this immune factor interacts with various bacterial and viral vaccines.

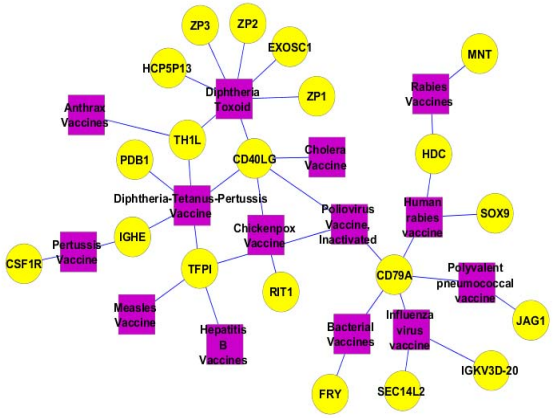

Fig. 4. A vaccine-gene subnetwork. The associations between vaccines and related genes were visualized by the Cytoscape tool (Smoot, Ono et al. 2011). Purple rectangular node represents vaccine, and yellow circle node represents gene.

4 DISCUSSIONS AND FUTURE WORK

In this paper, we proposed a novel network-based approach to investigate the vaccine relationships in the context of vaccine network extracted from PubMed literature abstracts. The investigations of vaccine-vaccine, vaccine-disease, and vaccine-gene networks demonstrate that such literature-based associations can be better analyzed using VO and such a combinatorial analysis is able to reveal the association patterns and identify new knowledge. The identified vaccine-vaccine associations based on vaccine-disease distance analysis are consistent with their VO categories and often lead to the generation of new hypotheses. Our studies discovered some novel vaccine-vaccine relationships by discovering a group of vaccines associated with some common diseases as demonstrated in the heat map analysis in the Results section. Due to the incompleteness of such information existing in the literature abstracts, such vaccine-vaccine associations need further validation in independent databases or through future experimental studies. For example, while our analysis reveals associations between a group of vaccines and neurological adverse events, it is noted that the evidences of these associations, although reported by some PubMed abstracts, are not necessarily commonly acknowledged (Sarntivijai, Xiang et al. 2012). More analysis may be required for clarification.

Future extensions of this work include: (1) integration of more comprehensive vaccine-disease association databases (e.g., VAERS system) to construct more complete vaccine-related networks; (2) generation of vaccine-related gene network by extending the neighbour genes of vaccine-associated genes; (3) network-based investigation of the relationships among vaccines and other drugs using vaccine-drug associations; (4) investigation on possible ways to improve the network by assigning weights or confident rates to different types of associations or associations from different sources.

ACKNOWLEDGEMENTS

This project was supported by the National Institute of Health grants 5R01LM009959-02 to HL, and R01AI081062 to YH.

REFERENCES

(2012). "The UMLS Semantic Groups." from http://semanticnetwork.nlm.nih.gov/SemGroups/.

Ashburn, T. T. and K. B. Thor (2004). "Drug repositioning: identifying and developing new uses for existing drugs." Nature reviews. Drug discovery **3**(8): 673-683.

Ball, R. and T. Botsis (2011). "Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS?" Clinical pharmacology and therapeutics **90**(2): 271-278.

Barabasi, A. L., N. Gulbahce, et al. (2011). "Network medicine: a network-based approach to human disease." Nature reviews. Genetics **12**(1): 56-68.

Berger, S. I. and R. Iyengar (2009). "Network analyses in systems pharmacology." Bioinformatics **25**(19): 2466-2472.

Bozzola, E., A. E. Tozzi, et al. (2012). "Neurological complications of varicella in childhood: case series and a systematic review of the literature." Vaccine **30**(39): 5785-5790.

Chong, C. R. and D. J. Sullivan, Jr. (2007). "New uses for old drugs." Nature **448**(7154): 645-646.

Comenge, Y. and M. Girard (2006). "Multiple sclerosis and hepatitis B vaccination: adding the credibility of molecular biology to an unusual level of clinical and epidemiological evidence." Medical hypotheses **66**(1): 84-86.

Corkins, M., C. Grose, et al. (1991). "Fatal pertussis in an Iowa infant." Iowa medicine : journal of the Iowa Medical Society **81**(9): 383-384.

Fekete, A., G. Vattay, et al. (2009). "Shortest path discovery of complex networks." Physical review. E, Statistical, nonlinear, and soft matter physics **79**(6 Pt 2): 065101.

Friedrich, F. (1998). "Neurologic complications associated with oral poliovirus vaccine and genomic variability of the vaccine strains after multiplication in humans." Acta virologica **42**(3): 187-194.

Goldstein, I., T. F. Lue, et al. (1998). "Oral sildenafil in the treatment of erectile dysfunction. Sildenafil Study Group." N Engl J Med **338**(20): 1397-1404.

Graul, A. I., L. Sorbera, et al. (2010). "The Year's New Drugs & Biologics - 2009." Drug news & perspectives **23**(1): 7-36.

Guess, M. J. and S. B. Wilson (2002). "Introduction to hierarchical clustering." Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society **19**(2): 144-151.

Hartung, H. P., B. Keller-Stanislawski, et al. (2012). "[Guillain-Barre syndrome after exposure to influenza]." Der Nervenarzt **83**(6): 714-730.

Hemachudha, T., D. E. Griffin, et al. (1988). "Immunologic studies of rabies vaccination-induced Guillain-Barre syndrome." Neurology **38**(3): 375-378.

Henderson, R., K. Oates, et al. (2004). "General practitioners' concerns about childhood immunisation and suggestions for improving professional support and vaccine uptake." Communicable disease and public health / PHLS **7**(4): 260-266.

Hur, J., A. Ozgur, et al. (2012). "Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining." Journal of biomedical semantics **3**(1): 18.

Ideker, T. and R. Sharan (2008). "Protein networks in disease." Genome research **18**(4): 644-652.

Kornbluth, R. S. (2000). "The emerging role of CD40 ligand in HIV infection." Journal of leukocyte biology **68**(3): 373-382.

Kornbluth, R. S. and G. W. Stone (2006). "Immunostimulatory combinations: designing the next generation of vaccine adjuvants." Journal of leukocyte biology **80**(5): 1084-1102.

Korsun, N., M. Kojouharova, et al. (2009). "Three cases of paralytic poliomyelitis associated with type 3 vaccine poliovirus strains in Bulgaria." Journal of medical virology **81**(9): 1661-1667.

Liu, Z., H. Fang, et al. (2013). "In silico drug repositioning: what we need to know." Drug discovery today **18**(3-4): 110-115.

Mathur, S. and D. Dinakarpandian (2011). "Drug repositioning using disease associated biological processes and network analysis of drug targets." AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium **2011**: 305-311.

Ozgur, A., T. Vu, et al. (2008). "Identifying gene-disease associations using centrality on a literature mined gene-interaction network." Bioinformatics **24**(13): i277-285.

Ozgur, A., Z. Xiang, et al. (2011). "Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology." Journal of biomedical semantics **2 Suppl 2**: S8.

Prevention, C. f. D. C. a. (1994). "Reported vaccine-preventable diseases--United States, 1993, and the childhood immunization initiative." MMWR. Morbidity and mortality weekly report **43**(4): 57-60.

Rindflesch, T. C., H. Kilicoglu, et al. (2011). "Semantic MEDLINE: An advanced information management application for biomedicine." Information Services & Use **31**(1/2): 15-21.

Sarntivijai, S., Z. Xiang, et al. (2012). "Ontology-based combinatorial comparative analysis of adverse events associated with killed and live influenza vaccines." PLoS One **7**(11): e49941.

Singhal, S., J. Mehta, et al. (1999). "Antitumor activity of thalidomide in refractory multiple myeloma." N Engl J Med **341**(21): 1565-1571.

Smoot, M. E., K. Ono, et al. (2011). "Cytoscape 2.8: new features for data integration and network visualization." Bioinformatics **27**(3): 431-432.

Tao, C., Y. Zhang, et al. (2012). Optimizing semantic MEDLINE for translational science studies using semantic web technologies. Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems. Maui, Hawaii, USA, ACM**:** 53-58.

Wardlaw, A. C. (1988). "Animal models for pertussis vaccine neurotoxicity." The Tokai journal of experimental and clinical medicine **13 Suppl**: 171-175.

Xiang, Z., M. Courtot, et al. (2010). "OntoFox: web-based support for ontology reuse." BMC research notes **3**: 175.

Building a Drug Ontology based on RxNorm and Other Sources

Josh Hanna,* Eric Joseph, Mathias Brochhausen, and William R. Hogan
Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

ABSTRACT

We built the Drug Ontology (DrOn) to meet the requirements of our comparative-effectiveness research use case, because existing artifacts had flaws too fundamental and numerous to meet them. However, one of the obstacles we faced when creating the Drug Ontology (DrOn) was the difficulty in reusing drug information from existing sources. The primary external source we have used at this stage in DrOn's development is RxNorm, a standard drug terminology curated by the National Library of Medicine (NLM). To build DrOn, we (1) mined data from historical releases of RxNorm and (2) mapped many RxNorm entities to Chemical Entities of Biological Interest (ChEBI) classes, pulling relevant information from ChEBI while doing so.

We built DrOn in a modular fashion to facilitate simpler extension and development of the ontology and to allow reasoning and construction to scale. Classes derived from each source are serialized in separate modules. For example, the classes in DrOn that are programmatically derived from RxNorm stored in a separate module and subsumed by classes in a manually built, realist, upper-level module of DrOn with terms such as 'clinical drug role', 'tablet', 'capsule', etc.

1 INTRODUCTION

An ontology of drugs could be useful for a variety of purposes, such as comparative effectiveness research (Olsen, 2011), clinical decision support (Broverman, 1998; Sperzel, 1998; Kim, 2001), and clinical data warehousing and integration (Broverman, 1998; Nelson, 2011; Palchuk, 2010; Parrish, 2006; Kim, 2001), among others. At present, no existing resource was sufficient for our use cases in these domains (see our sister paper (Hogan, 2013)), and therefore we decided to build the Drug Ontology (DrOn). Minimally, no existing resource contains in its current version a historically comprehensive list of National Drug Codes (NDCs).

RxNorm (Nelson, 2011)—a standard drug terminology maintained by the U.S. National Library of Medicine (NLM)—includes normalized names and relationships extracted from several proprietary drug knowledge bases. Because of (1) the large amount of drug information maintained within RxNorm, (2) the fact that it is freely available, and (3) the fact that much of it is available under a permissive license, RxNorm is a good candidate for a source of information to create a formal drug ontology.

RxNorm is focused primarily on prescription and over-the-counter drugs that are currently available in the United States. It uses Concept Unique Identifiers called RXCUIs to catalog and relate information.

At this stage of DrOn development, we are interested in the ability to query for National Drug Codes (NDCs). The NDC is a unique identifier that the Drug Listing Act of 1972 requires companies to report to the Food and Drug Administration (FDA). RxNorm associates each NDC with a drug product via the RXCUI. Although our requirement is to have a comprehensive, historical list of NDCs, RxNorm maintains only currently active NDCs in its current release. So tracking all NDCs and the RXCUIs with which they have been associated over historical releases of RxNorm is key to building DrOn.

Moreover, NDCs are often lost with no explanation when an RXCUI is retired, especially in releases of RxNorm prior to 2009. This situation necessitates careful tracking to ensure that all valid NDCs (and, indeed, any useful information) associated with a retired RXCUI can be associated with the most recent RXCUI that refers to the same entity.

While other drug information sources exist, none of them was sufficient. Our requirements include (1) a historically comprehensive list of NDCs, (2) correctness with respect to pharmacy and biomedical science, (3) logically consistent, correct axioms that do not entail untrue or inconsistent inferences, and (4) interoperability with other ontologies for translational science. In previous work, we analyzed RxNorm, the National Drug File – Reference Terminology, SNOMED CT, Chemical Entities of Biological Interest (ChEBI), an OWL conversion of the Anatomical and Therapeutic Chemical classification system, DrugBank, PharmGKB, and other sources (Hogan, 2013) and found that none of them met these requirements.

In this paper, we describe how we build DrOn from historical releases of RxNorm, while navigating these pitfalls. In addition, during the build process, we map drug ingredients from RxNorm to the Chemical Entities of Biological Interest (ChEBI) ontology (de Matos, 2010). As a result, we import hundreds of ChEBI classes and their associated URIs, labels, etc. into DrOn.

2 METHODS

The overall workflow of the extraction and translation process has three main steps:

1. Mining RxNorm for relevant entities, including information found only in older releases.
2. Extracting, Transforming, and Loading (ETL) the data mined from RxNorm into a normalized Relational Database Management System (RDBMS).
3. Translating the normalized RDBMS into an OWL 2.0 artifact.

* To whom correspondence should be addressed: jhanna@uams.edu

Each of these three steps is further subdivided into substeps that we explain in detail below.

2.1 Mining RxNorm

We first download the raw RxNorm data files directly from the NLM website, specifically the UMLS (or Unified Medical Language System) Terminology Services (UTS) site (U.S. National Library of Medicine, 2011) and import them into a locally hosted RDBMS using the scripts provided by the NLM. Additionally, to support maintenance of comprehensive information over time, we created and maintain two additional tables that store all the information that we extract from each release of RxNorm (a subset of all the information). We describe these tables in detail below (sections 2.1.3 and 2.1.4).

Currently, we include information from every version of RxNorm released between June, 2008 and February, 2013 in DrOn. The reason is that the June, 2008 release was the first one that includes RxNorm-curated NDCs.

It should be noted that we use only information curated within RxNorm and not any information from its sources directly, and thus our overall process is allowable under the UMLS license (all content reused in DrOn is Level 0).

2.1.1 RxNorm Files

The next step is to extract all relevant information from the files downloaded from the UTS site. RxNorm comes as a set of nine Rich Release Format (RRF) files, each of which contains a specific subset of the total information. However, we do not use all nine files in our build process.

We process RXNSAT.RRF, RXNCONSO.RRF, RXNCUI.RRF, and RXNCUICHANGES.RRF, RXNSAB.RRF. Table 1 shows the information we mined from each file.

File	Extracted Information
RXNSAT.RRF	NDCs and RXCUIs
RXNCONSO.RRF	RXCUI attributes
RXNCUI.RRF	retired RXCUIs with provenance
RXNCUICHANGES.RRF	RXCUI provenance
RXNSAB.RRF	RxNorm version information

***Table 1**: The RxNorm files and the information mined from each.*

There are four different term types in RXNCUI.RRF that are relevant to DrOn. They are: (1) Semantic Clinical Drug Forms (SCDFs), (2) Semantic Clinical Drugs (SCDs), (3) Semantic Branded Drugs (SBDs), and (4) Ingredients (IN). RxNorm treats NDCs as attributes of an SCD or SBD rather than a separate term type.

2.1.2 RXCUI Provenance

Tracking entities within RxNorm requires tracking the RXCUIs to which they are attached. This can be a difficult task. Any RXCUIs that have been entered in error are retired. Additionally, if two RXCUIs refer to the same entity, they are consolidated and either one of them is retired while the other remains or a new RXCUI is created and both older RXCUIs are retired. Prior to the April 2009 release of RxNorm, no comprehensive list of retired RXCUIs was provided. Furthermore, the reasons for retirement are not always well-documented, making it difficult to distinguish between RXCUIs that have been retired because they are nonsense and ones that have been replaced or merged. For instance, as of this writing, there are 40 RXCUIs with 210 associated NDCs that are no longer contained in the most recent release of RxNorm, however, there is no record of why these RXCUIs were removed.

2.1.3 Extraction of National Drug Codes (NDCs) and related RXCUIs

To facilitate the tracking of NDCs, we have created an additional table, *NDC_COMP*, that contains a comprehensive list of all RxNorm-curated NDCs from all releases of RxNorm since June 2008 (when they first appeared) and their corresponding RXCUIs. To generate this table, we parse the RXNSAT.RRF data file contained in each release of RxNorm. Any entry in the file whose source is RxNorm and is annotated as being an NDC is extracted from the file, along with its associated RXCUI, and imported into our *NDC_COMP* table. We also store the version from which each NDC was mined, which is parsed from the RXNSAB.RRF data file as mentioned in Section 2.1.1.

2.1.4 Tracking Provenance

The second of the two additional tables is a master conversion table, *DEPRECATED_RXCUIS,* which we use to track the current status of each retired RXCUI. This table contains two fields: *old_rxcui* and *new_rxcui.* The *old_rxcui* field contains a retired RXCUI, and the *new_rxcui* field contains the current RXCUI to which the retired RXCUI's information is now associated. The *new_rxcui* field may also contain a status code if the retired RXCUI's information is unable to be tracked because it was entered in error or split into multiple new RXCUIs. These special status codes are "ERROR" for RXCUIs that have been entered in error and "S_RXNCUI" for RXCUIs which have been split. Because RxNorm does not document why an erroneous RXCUI was entered in error, we are unable to do further processing on them or their associated information. For the RXCUIs which are split, it may be possible to track some of their associated information, but it is not always clear which in-

formation belongs to which child RXCUI and this issue requires manual intervention at present.

Our *DEPRECATED_RXCUIS* table is updated with each release of RxNorm through the following procedure:

1. First, we extract any RXCUIs from the comprehensive *NDC_COMP* table, built in Section 2.1.3, that can no longer be found in the *RXNCONSO.RRF* file being imported. We then import these RXCUIs into the *old_rxcui* column of our *DEPRECATED_RXCUIS* table. Because the *RXNCONSO.RRF* file contains all current RXCUIs, any RXCUIs that meet the above criteria must have been retired.
2. Next, using the RxNorm-curated *RXNCUI* table, we update all entries in the *new_rxcui* column. The RXNCUI table contains a *cui1* field containing a retired RXCUI, a *cui2* field containing the RXCUI into which the retired RXCUI's information has been merged, and a *cardinality* column contains the number of RXCUIs into which the information has been merged. Any RXCUI that has been entered in error is indicated by an entry in which the value of the *cui1* field is equal to the value of the *cui2* field. Additionally, any entry with a *cardinality* greater than 1 indicates that the RXCUI has been split. These are indicated in our table by setting the *new_rxcui* entry to "ERROR" and "S_RXNCUI", respectively. As of this writing, 768 RXCUIs and 3,484 associated NDCs are reported by RxNorm to have been entered in error and are therefore not included in DrOn. Additionally, 187 RXCUIs and 3,126 associated NDCs have been split. Both these RXCUIs and NDCs have also been left out of DrOn (for the time being) due to the difficulty of determining which information from the parent RXCUI belongs to which child RXCUI.
3. Finally, we compute the transitive closure, associating each RXCUI with the latest RXCUI that refers to the same entity with no intervening steps in our *DEPRECATED_RXCUIS* table. Because this table is updated with each release of RxNorm, occasionally an RXCUI in the *new_rxcui* field is retired. In such situations, the *new_rxcui* field is updated as described in Step 2, and a new row in the table is created with the newly-retired RXCUI set as the *old_rxcui*, and the *new_rxcui* field is set to match the updated *new_rxcui* from the original entry.

2.2 Mapping to ChEBI

The process maps ingredients (IN entity type) extracted from RxNorm to ChEBI entities where possible. We accomplish this step through a simple Java console application (that we built) that compares the labels of ingredients pulled from RxNorm with annotations in ChEBI. Any exact matches between the names or synonyms of RxNorm IN entities and ChEBI annotations were assumed to be referring to the same entity type and thus the ChEBI URIs were used in DrOn. Three different annotation types from ChEBI are used in the mapping process: *label, related_synonym*, and *exact_synonym*. To date, we import into DrOn ~750 classes (including URI and label and other annotations) from ChEBI: roughly 500 matches were found via *label*, 250 were found via *related_synonym*, and only two were found via *exact_synonym*. Many of the ingredients found in RxNorm are extracts of various plants, *e.g. ginger extract*, which we would not expect to find in ChEBI.

Somatropin (also known as somatotroin or human growth hormone) was erroneously associated with the ChEBI role 'growth hormone'. This error, once noticed, was fixed. The term is now mapped to the Protein Ontology URI that represents the protein molecule somatotropin.

We assigned a DrOn URI to every ingredient that was not found in ChEBI via this process.

2.3 ETL into a Normalized Format

There are five RxNorm entity types we were initially interested in pulling from RxNorm. These are the following: ingredient, clinical drug form, clinical drug, branded drug, and national drug code (NDC). Additionally, we wanted to represent a number of ingredient dispositions. Figure 1 shows these six entity types and the relationships between them. They are described in some detail next. It should be noted that the entities the NDC class represent are not the codes themselves, but instead the packaged drug products that the NDCs represent. Additionally, every DrOn entity that corresponds to a RxNorm entity is annotated with the corresponding RXCUI.

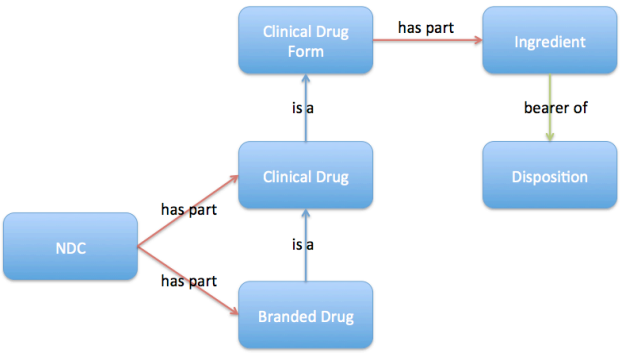

***Fig. 1**: The entity types of DrOn and their relationships as stored in the normalized format*

2.2.1 Entity types

The **ingredient** entities represent the types of molecules that are present in a drug product and have an active biological role. The URIs of ingredients, where possible, are taken from the Chemical Entities of Biological Interest (ChEBI) ontology as described above. Examples of these include

acetaminophen, sulfur, and ephedrine. There are 7,848 unique ingredients in DrOn.

The **disposition** entities represent dispositions that molecules bear (see Hogan, 2013). There are, as of now, six molecular dispositions in DrOn. They are:

1. *non-activating competitive beta-adrenergic receptor binding disposition* (i.e., beta-adrenergic blockade)
2. *function-inhibiting hydrogen/potassium adenosine triphosphatase enzyme (H+/K+ ATPase) binding disposition* (i.e., proton-pump inhibition)
3. *function-inhibiting L-type voltage-gated calcium channel binding disposition* (i.e., a subtype of calcium-channel blockade)
4. *function-inhibiting vitamin K epoxide reductase binding disposition* (i.e., a type of Vitamin K antagonism)
5. *function-inhibiting Na-K-Cl cotransporter 2 (NKCC2) binding disposition* (i.e., NKCC2 inhibition)
6. *function-inhibiting T-type calcium channel binding disposition* (i.e., another subtype of calcium-channel blockade)

These six dispositions were chosen based on their biological importance and relevance to ongoing comparative effectiveness research at the University of Arkansas for Medical Sciences. There is no direct correspondence between DrOn dispositions and RxNorm, because by design RxNorm lacks information about mechanism of action. Instead, the relationships between DrOn dispositions and ingredients was mined from ChEBI, although ChEBI treats the same realizable entities that we represent here as roles (see Hogan, 2013 for more details). Table 2 shows the associated ChEBI role from which the ingredient relationships for the three dispositions were mined. The other three dispositions not in the table were curated manually by the authors.

DrOn Disposition	ChEBI Role
non-activating competitive beta-adrenergic receptor binding disposition	beta-adrenergic antagonist
function-inhibiting hydrogen/potassium adenosine triphosphatase enzyme (H+/K+ ATPase) binding disposition	proton pump inhibitor
function-inhibiting L-type voltage-gated calcium channel binding disposition	calcium channel blocker

***Table 2**: The ChEBI roles used to mine DrOn disposition-ingredient relationships*

Function-inhibiting T-type calcium channel binding disposition was included because we erroneously associated *ethosuximide* and *function-inhibiting L-type voltage-gated calcium channel binding disposition*. This error was not due to any particular oversight of ChEBI but an artifact caused by the more specific nature of DrOn's dispositions as compared to ChEBI's more general c*alcium channel blocker*.

The **Clinical Drug Form** (CDF) entities represent a type of drug product, dose form (*e.g.* drug tablet), and, often, the intended route of administration (*e.g.* oral ingestion) without brand or strength information. These correspond to SCDFs in RxNorm. Examples of CDFs include *estradiol transdermal patch, iodine topical solution,* and *menthol crystals*. There are 14,035 unique CDFs in DrOn.

The **Clinical Drug** (CD) entities represent drug products with specific dosage/strength/form information. They are related to the CDF by an is-a relationship. For example, every *aspirin 325 MG enteric coated tablet* (CD) is a *aspirin enteric coated tablet* (CDF). DrOn contains 34,560 CDs.

The **Branded Drug** (BD) entities represent brand-name drug products with specific dosage/strength/form information. The drug products that BDs represent are related to the products that CDs represent by an is-a relationship. There are 21,248 unique BDs in DrOn.

The **National Drug Code** (NDC) entities represent a drug product and its packaging. These entities are distinct from entities represented by BDs or CDs, instead containing some number of instances of drug products represented by CDs/BDs, for example a 100-tablet bottle of aspirin 325 mg tablets. There are 390,813 unique NDC entities in DrOn.

DrOn Entity Type	RxNorm Entity Type
CDF	SCDF
CD	SCD
BD	SBD
Ingredient	IN
NDC	SCD or SBD Attribute

***Table 3**: The associated RxNorm entity type for each DrOn entity type except disposition.*

2.3.2 RDBMS design

The RDBMS design representing the normalized format of the entity types described above is simple. There are 5 core tables, one for each entity type. These are as follows: *clinical_drug_form, clinical_drug, branded_drug, ndc, ingredient,* and *disposition.*

Additionally, there are two tables storing provenance information from RxNorm, such as the version of RxNorm in which each RxCUI was found. These are *rxcui* and *rxnorm*. These are completely separate from the core entity tables to allow for incorporation of other data.

Many-to-many tables representing the relationships between the various entities are omitted in the interest of brevity. However, all of the relationships shown in 1 are also represented in RDBMS.

2.3.2 ETL process

The ETL process is done in four major steps:

1. First, we initialize the rxcui and rxnorm tables. This includes mapping every deprecated RXCUI to the most recent RXCUI that identifies the same object, either to an RXCUI from the current set or another deprecated, but not entered in error, RXCUI.
2. Next, we initialize the ndc table. This primarily involves copying all the NDCs found in the mining process (without the duplication caused by storing NDCs multiple times during the mining process) and associating them with the relevant RXCUI.
3. Next, we create the ingredients, CDFs, CDs, and BDs from the associated RxNorm type. This includes maintaining the proper relationships between the various entities (e.g. associating the correct ingredients with each CDF).
4. Finally, we associate each NDC with the appropriate CD or BD. This primarily involves following the provenance trail of RXCUIs provided in step 1.

2.4 Creating the OWL 2.0 Artifact

We use the OWLAPI 3.4.3 (Horridge, 2011), Scala 2.10 (Odersky, 2004), and Slick 1.0.0 (Typesafe, 2013) to extract the entities from our internal representation and transform them into an OWL artifact. This process is subdivided into the following steps:

1. Extract the ingredients, using ChEBI URIs where appropriate.
2. Extract the dispositions and associated them via the *bearer_of* relation to the one or more ingredients.
3. Extract the clinical drug forms and associate them via the *has_proper_part* relation to the one or more ingredients.
4. Extract the clinical drugs and assert they are a subclass of the appropriate clinical drug form.
5. Extract the branded drugs and assert that they are a subclass of the appropriate clinical drug.
6. Extract the NDCs and assert that they are related to one branded drug or one clinical drug via the *has_proper_part* relation.

This ordering of the steps is deliberate. Each step depends on one or more previous steps.

Since the RDBMS structure defined above represents the entities and their relationships already, this process is fairly straightforward.

2.4.1 Modularization

The ability to incorporate additional sources of information has been a key requirement for the build process. To help facilitate this, we developed DrOn in a modular fashion. Currently, DrOn has five different modules: **dron-full**, **dron-chebi**, **dron-rxnorm**, **dron-pro**, and **dron-upper**.

The **dron-full** module is simply a connector that imports the other modules. It is so named on the assumption that certain subsets of the modules may prove useful enough to warrant lighter versions of the ontology.

The **dron-chebi** module contains all of the annotations for the ingredients mapped to ChEBI (as described in Section 2.2). It contains everything imported from ChEBI.

The **dron-rxnorm** module contains all of the information mined from RxNorm, which, at this point of the ontology's development, is the bulk of DrOn's information. It includes the NDCs, though we plan to split the NDCs from the rest of the RxNorm module in future work.

The **dron-pro** module includes everything imported from the Protein Ontology (PRO). At present, it is very small and only contains the 'protein' and 'somatotropin' classes from PRO. As stated above, we imported these classes to represent somatotropin as a drug ingredient.

The **dron-upper** module contains the hand-created upper level ontology that the other modules are mapped on to (see Hogan, 2013).

This modularization brings two major benefits: development simplicity and increased scalability. By creating logical divisions and well-defined interfaces between the modules, we can more easily maintain each module separately without significantly affecting the other modules. Additionally, as each module grows in size, we can shard the processing and creation of the ontologies to different servers, making scaling the process simpler.

3 DISCUSSION

We developed an ontology, DrOn, that contains information programmatically derived from three different sources (RxNorm, ChEBI, and PRO) during its build process. Because it is derived from general-purpose resources, we believe DrOn can serve many use cases beyond our current ones (although this conjecture requires further research). We plan on adding additional sources in the future to maintain current information in DrOn, with more immediate plans to include information from Structured Product Labels. As such, we built our internal representation to maintain provenance information of the sources separately, ensuring that we can both track the provenance of the various entities as the ontology develops and add new sources without adversely affecting the existing ontology.

DrOn follows OBO Foundry guidelines and is currently listed on the OBO Foundry website as a candidate ontology. In additional to the mining detailed above, DrOn imports BFO 1.1 and includes terms MIREOTed from the Relationship Ontology and BFO 2.

The development site and issue tracker for DrOn can be found at https://bitbucket.org/uamsdbmi/dron. The perma-

nent URL for DrOn is http://purl.obolibrary.org/obo/dron.owl.

Our primary, driving use case was support for Comparative Effective Research. Author WRH was part of a research team wherein a student had to manually identify all drug products that contain acetaminophen historically. We built a web application that uses DrOn to support this use case; users can search for all NDCs that either contain a specific ingredient or contain an ingredient that realizes a specific disposition. This web application is accessible at http://ingarden.uams.edu/ingredients.

Future work includes addressing limitations in the current process. One of the more egregious examples is the lack of a link from the various drug products to their dose forms (*e.g.,* drug capsule). Nearly all of the most common dose forms are already in the upper level of the ontology (**dron-upper** module), but the CDFs are not properly related to them. This is due to (1) time constraints and (2) the dubious ontological nature of some of the dose forms found in RxNorm. For example, 'inhaler' does not refer to the form of the drug but instead to its container (which also serves the role of drug delivery device). But the form of the drug itself is a solution or suspension contained in the inhaler. Note that the presentation form in this case (e.g., solution) differs from the administration form (e.g., aerosol).

Another issue is the lack of a full logical definition for some of the terms. For instance, only a small subset of the parts of each drug product is defined. A *clinical drug form* contains information about its dose form, its route of administration, and its active ingredients. As of the writing of this paper, the only one of these that is represented in the ontology as classes are the active ingredients, though dose forms are mostly represented. Even these, however, are still not fully developed, generally lacking any class restrictions. Additionally, a *clinical drug* contains dosage information and *branded drugs* have brand information. Neither of these is represented in the ontology.

The final issue with the process is the need for manual interaction. Although each step within in the process is automated, they are not tied together in a coherent way. We expect that some manual intervention will always be needed as we continue to mine updated information from these sources, but there is significant room for improvement in connecting the various segments of the overall process flow and fully automating the less ontologically nebulous steps.

Since DrOn is already large, and will likely increase in size as we incorporate more sources and as more drug products are manufactured, we expect that we will run into difficulties managing generation of and reasoning over the ontology. One potential solution we intend to investigate is to reason over the modules individually and combine the results. We also intend to create more manageable subsets of DrOn, which should allow users to work with only the portions of DrOn that they need for a particular use case.

ACKNOWLEDGEMENTS

This work was supported by award number UL1TR000039 from the National Center for Advancing Translational Sciences, award R01GM101151 from the National Institute for General Medical Science, and the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000. This paper does not represent the views of NCATS, NIGMS, or NIH.

REFERENCES

Nelson, S.J., Zeng, K., Kilbourn, J., Powell, T., &Moore, R. (2011). Normalizied names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association: JAMIA, 18(4), 441 - 448

U.S. National Library of Medicine (2011). RxNorm Retrieved April 24, 2013, from http://www.nlm.nih.gov/research/umls/rxnorm/

de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., and Steinbeck, C. (2010). Chemical Entities of Biological Interest: an update. *Nucl. Acids Res.*, **38**, D249–D254.

Horridge, M., Bechhofer, S. (2011). The OWL API: A Java API for OWL Ontologies. *Semantic Web Journal*, 2(1), 11 – 21

Odersky, M., Altherr, P., Cremet, V., Dragos, I., Dubochet, G., Emir, B., McDirmid, S., Micheloud, S., Mihaylov, N., Schinz, M., Stenman, E., Spoon, L., Zenger, M. (2004). An Overview of the Scala Programming Language, *Technical Report LAMP-REPORT-2006-001*

Typesafe (2013), Slick Retrieved April 24, 2013, from http://slick.typesafe.com/

Broverman, C., Kapusnik-Uner, J., Shalaby, J., & Sperzel, D. (1998). A concept-based medication vocabulary: an essential requirement for pharmacy decision support. Pharmacy practice management quarterly, 18(1), 1-20.

Palchuk, M. B., Klumpenaar, M., Jatkar, T., Zottola, R. J., Adams, W. G., & Abend, A. H. (2010). Enabling Hierarchical View of RxNorm with NDF-RT Drug Classes. AMIA Annual Symposium proceedings, 2010, 577-581.

Parrish, F., Do, N., Bouhaddou, O., & Warnekar, P. (2006). Implementation of RxNorm as a Terminology Mediation Standard for Exchanging Pharmacy Medication between Federal Agencies. AMIA Annu Symp Proc, 1057.

Olsen, L., Grossman, C., & McGinnis, J. M. (2011). Learning What Works: Infrastructure Required for Comparative Effectiveness Research: Workshop Summary: The National Academies Press.

Sperzel, W. D., Broverman, C. A., Kapusnik-Uner, J. E., & Schlesinger, J. M. (1998). The need for a concept-based medication vocabulary as an enabling infrastructure in health informatics. Proceedings AMIA Annual Symposium. 865-869.

Kim, J. M., & Frosdick, P. (2001). Description of a drug hierarchy in a concept-based reference terminology. Proc AMIA Symp, 314-318.

Hogan, W. R., Hanna, J., Joseph, E., Brochhausen, M. (2013). Towards a Consistent and Scientifically Accurate Drug Ontology. ICBO 2013 Conference Proceedings

Maintaining the Drug Ontology: an Open-source, Structured Product Label API for the JVM

Roger A. Hall*, Josh Hanna, and William R. Hogan
Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

ABSTRACT

Our use case for maintenance of the Drug Ontology includes a semi-automated, daily process capable of importing new, relevant information from a variety of linkable resources, using fast and flexible algorithms with full access to all data. Structured Product Labels contain linkable information regarding FDA approved drug products and the drug packages in which they are sold, as well as ingredients and metadata about the drugs. We created an Application Programming Interface for SPLs using Scala, which will run on any implementation of the Java Virtual Machine (JVM) and is freely available through an open-source license for any non-commercial use.

1 INTRODUCTION

We are using Structured Product Labels (SPL) from the Food and Drug Administration (FDA) to capture new drug and related entities for representation in the Drug Ontology (DrOn). The reason is that manufacturers must submit an SPL for all new drugs approved in the United States, and thus this information is "directly from the source". It is also much richer than what is available in drug terminologies such as RxNorm. This paper describes our processes and the software we are developing to support them for extracting information from SPLs to update DrOn as new drug products come into existence.

SPLs are machine-readable files, submitted-to and released-by the FDA, containing prescription drug labeling and product metadata, such as National Drug Codes (NDCs) and drug product ingredients. They are available as full "current revision" releases, and monthly, weekly, or daily updates from the DailyMed website at http://dailymed.nlm.nih.gov. Each archive release contains individual archives. Each individual archive contains one XML file and may contain zero or more ".jpg" image files, which are referenced by the XML file.

SPLs have been found useful linking active ingredients and chemical entities (Hassanzadeh *et al*., 2013), extracting indication information (Fung *et al*., 2013), and improving detection of drug-intolerance issues (Schadow, 2009) and can be enhanced with current literature for greater safety, efficacy, and effectiveness (Boyce *et al*., 2013).

While a non-proprietary SPL parsing web service called "LinkedSPLs" is available (Hassanzadeh *et al*., 2013), we discuss below the lack of fitness for our use case.

Recent work (Hogan *et al*., 2013) has shown the benefit of ontological realism for avoiding scientific inaccuracy with the creation of the Drug Ontology (DrOn). In addition to modeling drug products, ingredients, and their respective dispositions, DrOn includes an extensive historical collection of identifiers such as NDCs. To keep DrOn updated as new drug products come into existence, we have designed a system for automated staging of data from three initial sources: RxNorm, ChEBI, and SPLs (Fig 1).

In addition to the need to drive DrOn maintenance, we are aware that SPL files are created by a large and diverse user base in industry and submitted to the FDA, so we have also included the capability to write SPL documents. Although

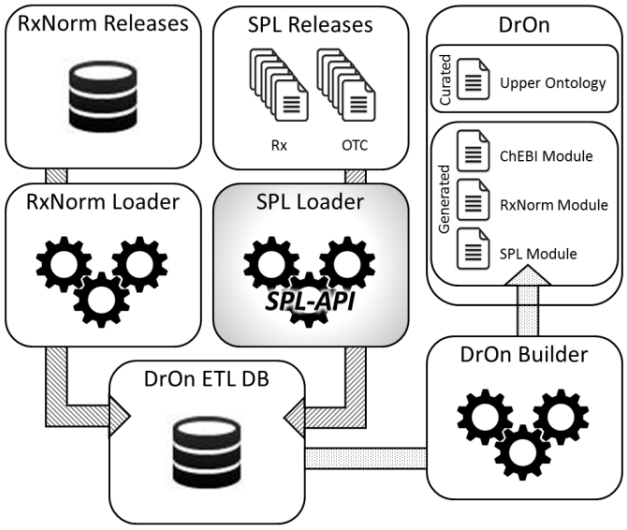

Figure 1: SPL-API is a part of the DrOn ETL process for automatically regenerating DrOn modules holding externally sourced information on drugs, ingredients, and dispositions.

one free tool for creating and editing SPL files exists ("SPL XForms"), we discuss here our motivations for creating a new tool.

Our implementation is an Application Programming Interface (API) for SPLs (SPL-API) which we use to update DrOn each day using the "daily updates" from Dailymed.

* To whom correspondence should be addressed: rahall2@uams.edu

SPL-API also includes features for downloading and parsing full releases and monthly and weekly updates.

Thus, the goal of this work is to create a generally useful open-source parser and writer for SPL XML files, and to use it within a larger system to update DrOn with applicable classes and their relationships, enabling additional data to be linked by other processes. As a result DrOn will be continuously updated with new drug products as they are approved.

2 METHODS

The DrOn support system must Extract, Transform, and Load (ETL) data from a variety of sources in order to automatically rebuild the "lower ontology" containing specific drug products from the latest sources. A DrOn Builder has been previously implemented (Hanna *et al*., 2013) which produces OWL 2.0 ontology modules from the initial DrOn database. Here, we describe the methods completed and planned to add SPL support to DrOn Builder.

2.1 Analysis of Existing SPL Tools

Our use case requires full access to all data in available resources and flexible server-side processing, preferring a local library over a connected service for both processing speed and algorithmic flexibility.

LinkedSPL provides SPL content for prescription and over-the-counter drugs, and is updated weekly. It can be accessed through a SPARQL endpoint to acquire the free-text contained in any "section" of the SPL file, which is defined by the "<section>" tag-set. Although the LinkedSPL software artifacts are freely available, and may be used locally, they are unable to report included label image files or a link to them. LinkedSPL also only parses prescription and OTC files, leaving out the "Remainder" labels (which include data on vaccines and some medical devices) and the "Animal" labels. Additionally, DrOn is not currently using RDF technologies (other than serialization of OWL into RDF), so we seek to avoid the complexity of adding an intermediate representation to the system.

A browser-based editor ("SPL XForms") for SPL format XML files is also available (Pragmatic, 2010). Developed in collaboration with the FDA, it can be used to view, create, edit, and validate the XML data once a Java-helper is allowed to load. Although useful to our study of individual XML files, it is not freely available as a local library, and thus could not be part of future system integrations.

2.2 Analysis of SPL Labels

Software was written to survey all XML elements and their attributes and relationships. Survey data will be available online (see section 3). An analysis was conducted on 45,182 SPL submissions in the Human Prescription, Human Over The Counter, Medical Device, and Animal label sets available as of April 22, 2013. The survey revealed elements that were primarily classes, those that were primarily attributes, and those that were unnecessarily verbose "wrapper" elements. Additionally, elements which are found in collections were identified using a "max and mean" algorithm.

SPL Documents have a fairly simple structure (Fig 2), combining a metadata-filled header and a body (contained in element <structuredBody>). The body contains a list of "section" elements. Section elements contain other section elements. While 90% of all files had 24 levels of nesting or less, some runaways include 40 levels. We note that every element deeper than 18 levels is related to a nested <containerPackagedProduct> element, which creates significant ambiguity for parsing drug products.

Each section is "typed" by the *loinc_code* attribute according to LOINC codes (e.g. "34067-9") that identify the common sections of SPLs (e.g. "Indications and Usage"). There are 87 codes allowed per ucm162057.htm (FDA, 2013), but only 84 were observed. Most documents include an SPL PRODUCT DATA ELEMENTS SECTION, an INDICATIONS & USAGE SECTION, and a WARNINGS SECTION. There was a mean of 1.48 PACKAGE LABEL.PRINCIPAL DISPLAY PANEL sections per document. All other codes were observed in less than half of the documents (and most were observed in less than 20% of the documents), while a full 33% of all SPL sections were coded as SPL UNCLASSIFIED SECTION, suggesting significant limitations in the standard.

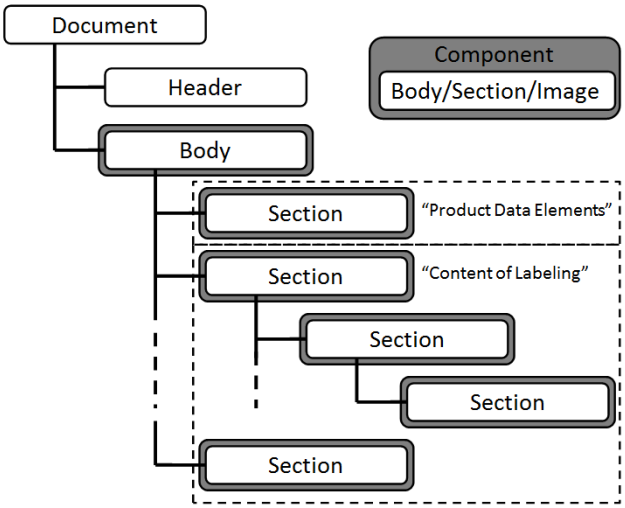

Figure 2: The SPL Document structure includes "wrapper classes" like <Component /> and "infinitely nest-able" <Sections />.

2.3 Technical Specifications

Our implementation is in Scala (version 2.10.1), which runs on any Java Virtual Machine (JVM) implementation and can be used within custom Java or Scala programs.

2.4 The DrOn Relational Schema

DrOn is influenced by RxNorm, and contains OWL 2.0 classes that model ingredients, semantic clinical drug forms,

semantic clinical drugs, and semantic branded drugs (Hogan *et al*., 2013). The staged data from all external resources used to build the "lower ontology" in DrOn are stored in a relational database whose schema follows the RxNorm file format closely. We added additional tables to the core schema for annotations regarding provenance, including a system-standard field for an "external link" to a resource-specific table. The external link can be used as an ID to load a resource-specific helper module as described below. At a minimum, the ID enables provenance for the external resource file. A "module" of resource-specific tables may be added to capture desired data. Although persistence of the all SPL information is unnecessary for our current integration with DrOn, our implementation represents all XML classes and attributes (except for the lowest level classes that represent HTML *formatting* of product labels). An applicable prefix for the table-set (e.g. "*spl_*") helps separate the tables visually when added to the same database.

The database currently holds on ~10^6 entries; the authors are experienced with databases containing ~10^9 entries. In the short term, expansion of ingredients and dispositions will increase the database more quickly than new products.

2.5 XML to JVM Classes with Code Generation

Code generation (cogen) has been shown useful for creating packages with numerous classes from OWL ontologies (Kalyanpur *et al*., 2004). It has also been useful in generating SOAP clients from WSDL files (Simpkins, 2008).

We developed a custom code generation utility to generate Scala classes for the SPL XML Format using the referenced XML Schema Definition (XSD), which validates the SPL format, and the survey results (see section 2.2). Classes were identified as elements (and their wrappers) which contain a number of attributes and zero or more collections. Attributes and elements of collections may both be typed as other classes. Collections were implemented to hold lists of child node types when necessary. Node types that never contain other node types, such as <id />, <name>, and <code />, are created as typed attributes of the classes that represent the containing node types. Accessors were generated for attributes; iterators were generated for collections.

Instead of attempting to create classes at run-time through the Java-beans paradigm (Kalyanpur *et al*., 2004), we chose to keep code generation in a separate utility, and import the resulting "top 34" classes into the dron.spl package.

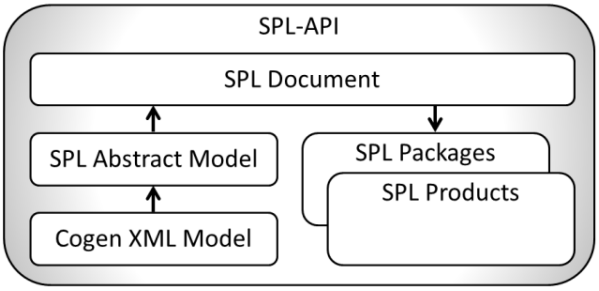

Figure 3: The SPL-API uses a set of 34 cogen classes that model the XML precisely within a set of classes that model an SPL document abstractly. The figure shows data flow for parsing.

2.6 The SPL-API

The parser and writer implementation is contained in the package *dron.spl*. Three sets of classes are included; classes that model the SPL XML format, classes that model the products and packages represented in the SPL XML files, and utilities necessary to manage Dailymed releases and updates (Fig 3 does not show utilities).

The root Scala class is "SPLDocument". At least one SPLDocument instance is created for each SPL submission file parsed (see section 2.8).

2.6.1 SPL XML Format

SPLDocument exposes classes and methods that represent an abstract SPL document, which in turn utilize the cogen classes that model the XML more exactly (Fig 3). A recursive printing algorithm built into the cogen classes enables the SPL-API to write SPL files.

When an XML element is always wrapped by another element, and the parent element never contains another element, then only one cogen class is represented. There are fourteen wrapped classes in SPL-API.

As an example of the layered class design, consider the "ComponentStructuredBody" cogen class that represents the XML elements for the SPL document body. (This class is the "structuredBody" element wrapped with a "component" element.) With this class, you can create a collection of "ComponentSection" cogen classes. However, the better approach would normally be to use the "section methods" of SPLDocument (e.g. *SPLDocument.addSection()*) to manage the sections of a document.

2.6.2 SPL Drug Classes

Classes for Drug Products, Ingredients, and Drug Label Data were created, along with a base class for Drug Packages. Ingredients are maintained as a collection in the Products class since each product may contain multiple ingredients, and each ingredient has "active" or "inactive" status attribute. Additional attributes are planned, such as the "strength" of the Ingredient within the Drug Product.

The primary subject of an SPL file—a drug package—is implemented as two classes that are sub-classed from the base class Package; SimplePackage and ComplexPackage. Every instance of Package must be related to at least one instance of Product. SimplePackage relates to exactly one NDC, while a ComplexPackages contains a collection of SimplePackage(s) along with its own metadata.

Parsing one SPL file produces a list of Package instances, which will have one or more elements of Content of Labeling Data. Lists of Label Data are maintained as a collection in the appropriate Package instance.

2.6.3 SPL-API File Utilities

We provide utilities for downloading full release and periodic updates, unzipping downloaded archives, unzipping all archives in a given directory, and unzipping individual submission archives. Additional data lookup utilities will be added, for example to translate *loinc_code*(s) to text labels, which is hoped to also assist users in minimizing the future share of SPL UNCLASSIFIED sections.

2.7 Matching NDCs

For our use case, a key step in correctly identifying all of the real drug products represented by the XML submission file is to identify all NDCs, but NDCs are not encoded in the XML scheme, and are only found in the free text of Product Data Elements sections. They generally contain the text string "NDC", and they always conform to the NDC 10-digit format (5-4-1, 4-4-2, or 5-3-2). We use pattern matching to identify multiple NDCs per text section. The NDCs that are found are checked against the National Drug Code Directory (FDA, 2013). The ability to correctly match all NDC's affects the quality of the results of the Drug Package listing (see section 2.6.2).

2.8 Core Document References

Of the 45,182 SPL submissions surveyed, 220 used the XML tag-set "<relatedDocument>". This tag includes the "SetID" of a "Core Document Reference" (FDA, 2012) (CDR), from which all sections are inherited by the containing document. When parsing a submission XML, the SPL-API will load a related document if it is available within the same directory, and create a separate instance of SPLDocument to hold the related document data. Documents that are "parents" can still have a <relatedDocument> tag, so the loading scheme is recursive, and is currently dependent on a small level of nesting.

When using the Scala classes that represent the XML model (see section 2.6.1), each SPL section is contained within its proper document, and each related document is accessible by the *SPLDocument.getRelatedDocument()* property accessor.

When using the Scala classes that represent Drug Packages (see section 2.6.2), all related documents are "flattened", and each section is included from all documents. Inheritance rules are unclear, so all sections are currently collected by section type. All identified Drug Products and Drug Packages will be included in the list.

2.9 ETL and External Resource Helpers

In addition to the SPL-API, a "helper" will be developed to load the parsed SPL data into the DrOn relational schema (represented as gears in Fig 1). A plugin system added to the DrOn builder will be able to identify the proper resource-specific plugin and pass the initialization necessary to complete loading for the next update.

3 CONCLUSION

We have developed an open-source API for processing SPLs in a Java Virtual Machine. A developer's release will be made available at the start of VDOS 2013, and will be available at: https://bitbucket.org/rogerhall68/spl-api.

Ongoing work includes loading processed data into the Drug Ontology to keep it current as new drug products are released.

ACKNOWLEDGEMENTS

This work was supported by award number UL1TR000039 from the National Center for Advancing Translational Sciences, award R01GM101151 from the National Institute for General Medical Science, and the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000. This paper does not represent the views of NCATS, NIGMS, or NIH.

REFERENCES

Dailymed, http://dailymed.nlm.nih.gov/dailymed/downloadLabels.cfm

Hassanzadeh, O., Zhu, Q., Freimuth, R., & Boyce, R. (2013) Extending the "Web of Drug Identity" with Knowledge Extracted from United States Product Labels. *Proceedings of the 2013 AMIA Summit on Translational Bioinformatics*

Fung K.W., Jao C.S., & Demner-Fushman D. (2013) Extracting drug indication information from structured product labels using natural language processing. J Am Med Inform Assoc. 2013 May 1;20(3):482-8

Schadow, G. (2009) Structured Product Labeling Improves Detection of Drug-Intolerance Issues. *J. Am Med Inform Assoc*, **16**, 211–219.

Boyce, R., Horn J.R., Hassanzadeh O., de Waard A., Schneider J., Luciano J.S., Rastegar-Mojarad M., and Liakata M. (2013) Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness. *J. Am Med Inform Assoc*, **16**, 211–219.

Hogan, W.R., Hanna, J., Joseph, E., and Brochausen, M. (2013). Towards a Consistent and Scientifically Accurate Drug Ontology, This Volume.

Hanna J., Brochausen M., & Hogan W. R. (2013) Building a Realist Drug Ontology using RxNorm and Other Sources. This Volume.

FDA (2013) ucm162057.htm http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm162057.htm

Pragmatic Data (2010) SPL XForms, http://pragmaticdata.com/spl/form/

Kalyanpur, A., Pastor, D. J., Battle, S., & Padget, J. (2004, June). Automatic mapping of OWL ontologies into Java. In *SEKE* (Vol. 4, pp. 98-103).

LinkedSPL, http://dbmi-icode-01.dbmi.pitt.edu/linkedSPLs/

SPL XForms, http://pragmaticdata.com/spl/form/

Simpkins N., Generating a client from WSDL, http://www.eclipse.org/webtools/community/education/web/t320/Generating_a_client_from_WSDL.pdf

FDA (2013) National Drug Code Directory http://www.fda.gov/drugs/informationondrugs/ucm142438.htm

FDA (2012) Structured Product Labeling (SPL) Implementation Guide with Validation Procedures http://www.fda.gov/downloads/ForIndustry/DataStandards/StructuredProductLabeling/UCM321876.pdf

The Ontology of Vaccine Adverse Events (OVAE) and its usage in representing and analyzing vaccine adverse events

Erica Marcos[1], Yongqun He[2*]
[1] College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI 48109, USA.
[2] Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, Center for Computational Medicine, and Comprehensive Cancer Center, University of Michigan, Ann Arbor, MI 48109, USA.

ABSTRACT

Licensed human vaccines can induce various adverse events in vaccinated patients. Many known vaccine adverse events (VAEs) have been recorded in the package inserts of commercial vaccine products. To better represent and analyse VAEs, we developed the Ontology of Vaccine Adverse Events (OVAE) as an extension of the Ontology of Adverse Events (OAE) and the Vaccine Ontology (VO). OVAE has been used to represent and classify the adverse events recorded in package insert documents of commercial vaccines licensed by the USA Food and Drug Administration (FDA). OVAE currently includes over 1100 terms, including 87 distinct types of VAEs associated with 63 human vaccines licensed in the USA. Specific VAE occurrence rates associated with different age groups have been recorded in OVAE. SPARQL scripts were developed to query and analyse the OVAE knowledge base data. The top 10 vaccines accompanying with the highest numbers of VAEs and the top 10 VAEs most frequently observed among vaccines were identified and analysed. Different VAE occurrences in different age groups were also analysed. The ontological representation and analysis of the VAE data associated with licensed human vaccines improves the classification and understanding of vaccine-specific VAEs which supports rational VAE prevention and treatment and benefits public health.

1 INTRODUCTION

Many licensed vaccines exist to protect against a variety of diseases and infections. They are extremely useful in decreasing infection prevalence in human populations. Due to the public health benefits of vaccines, their coverage has been increasing in recent years. Each vaccine often induces different types of adverse events. As vaccine usage increases, the risk of adverse events proportionally increases (Cunha, Dorea, Marques, & Leao, 2013). Many known vaccine adverse events (VAEs) have been recorded in the package inserts of commercial vaccine products. There is also a need to predict probabilities of different adverse events arising in different individuals, which can potentially lead to a decline in the risk of developing an adverse event.

Two existing ontologies are closely related to the VAE studies. The Ontology of Adverse Events (OAE) is a community-based biomedical ontology in the area of adverse events (He, Xiang, Sarntivijai, Toldo, & Ceusters, 2011; Sarntivijai et al., 2012). OAE defines an 'adverse event' as a pathological bodily process that occurs after a medical intervention (*e.g.*, vaccination, drug administration). The OAE 'adverse event' is a subclass of the ontology term 'pathological bodily process' defined in the Ontology of General Medicine Science (OGMS) (http://code.google.com/p/ogms/). To be consistent with most practice uses of the term, OAE does not assume a causal relation between an 'adverse event' and a medical intervention. OAE has defined over 2,000 types of adverse events that are commonly found in different medical interventions. The community-based Vaccine Ontology (VO) represents various vaccines, vaccine components, and vaccinations (He et al., 2009; Lin & He, 2012). Both OAE and VO are OBO Foundry candidate ontologies and are developed by following the OBO Foundry principles (Smith et al., 2007).

OAE has been shown to significantly increase the power of analysis of case report data from the Vaccine Adverse Event Reporting System (VAERS) (Sarntivijai, et al., 2012). However, there has been no published paper to analyze commonly known VAEs recorded in the package insert documents of FDA licensed vaccines. Compared to the often noisy data creating difficulty in identifying the causality, the adverse events recorded in the official package inserts are known adverse events to specific vaccines.

To better represent various VAEs and support vaccine safety study, we developed the Ontology of Vaccine Adverse Events (OVAE) as an extension of the biomedical ontologies OAE and VO. In this paper, we introduce the basic framework of the OVAE and how OVAE is used to represent and analyze all adverse events reported in the product package inserts of commercial vaccines currently used in the USA market.

2 METHODS

2.1 OVAE ontology generation

Following VO and OAE, OVAE is also edited with the Web Ontology Language (OWL2) format (http://www.w3.org/TR/owl-guide/). FDA-licensed human vaccines represented in VO were imported to OVAE using the tool OntoFox (Xiang, Courtot, Brinkman, Ruttenberg, & He, 2010). Those adverse event terms reported in the package inserts of FDA licensed human vaccines were also imported to the OVAE using OntoFox.

* To whom correspondence should be addressed:
yongqunh@med.umich.edu

New OVAE-specific terms were generated with IDs containing the prefix of "OGSF_" followed by seven auto-incremental digital numbers and edited using the Protégé 4.2 OWL ontology editor (http://protege.stanford.edu/).

2.2 Data source of known VAEs

The official FDA website that provides supporting documents of licensed vaccines was the primary data source (FDA, 2013). A PDF version of a package insert document is available for almost every vaccine in the data source. The PDF document includes a section called "Adverse Reactions" that contains text descriptions of known vaccine adverse events associated with the vaccinated population.

2.3 Data collection and formatting to ontology

Based on the OVAE framework and the adverse event description in the package inserts, a design pattern was first generated to lay out the relations between different ontology classes, properties, terms and data types. The design pattern was used to form an MS Excel template for collection of individual adverse events for different vaccines. All the data in each package insert were manually examined and input to the Excel worksheet. Following the manual data collection and annotation, the program Ontorat (http://ontorat.hegroup.org) was used to transform the Excel file data to the OVAE ontology format (Xiang, Lin, & He, 2012).

2.4 VAE data analysis

To identify specific OAE or VO hierarchical structure among a list of terms, OntoFox was first used to extract the input OAE or VO terms and all associated terms required for proper hierarchical assertion and inference. The output OWL files were then visualized using a Protégé OWL editor. The SPARQL Protocol and RDF Query Language (SPARQL) is a W3C recommended language to query OWL RDF tripe store ("SPARQL query language for RDF,"). After OVAE was also deposited in the Hegroup RDF triple store, SPARQL was used for querying the OVAE knowledgebase from the RDF triple store to address a list of scientific questions.

2.5 OVAE ontology websites and license

The OVAE source code is available in a Google Code website: http://code.google.com/p/ovae. The OVAE project website is: http://www.violinet.org/ovae. OVAE has been deposited to the BioPortal project of the National Center of Biomedical Ontology (NCBO) (http://bioportal.bioontology.org/ontologies/3227). OVAE is also deposited in the Ontobee linked data server (http://www.ontobee.org/browser/index.php?o=OVAE) (Xiang, Mungall, Ruttenberg, & He, 2011). The OVAE source is freely available under the Apache License 2.0.

3 RESULTS

3.1 OVAE system design and statistics

The goal of current OVAE development is to generate an ontology-based VAE knowledgebase that represents known adverse events (AEs) associated with licensed vaccines. Such a knowledgebase incorporates the OAE terms of AEs together with the vaccine information defined in the VO. As the primary developer of the OAE and VO, we argue that OAE is not appropriate or responsible for representing various AEs specific for any particular medical intervention including vaccination due to the following reasons. First, OAE emphasizes the representation of various AEs general for most medical interventions and related topics (*e.g.*, methods for analysis of the causal relation between AEs and medical interventions, and factors affecting the causality analysis). Currently OAE is already large and contains over 3,000 terms. It is expected that many more AE terms will be added to OAE. Therefore, it is ideal to make OAE focused and as concise as possible. Secondly, AE researchers related to specific medical intervention domains may have more domain-specific demands and requests. For example, VAE researchers would like to link AEs to different vaccines. The vaccine (or drug) researchers may not be interested in drug (or vaccine) specific AEs. As a relatively independent domain, VAEs have been focuses of many vaccine researchers and groups. Independent from drug AEs, clinical VAEs are reported to vaccine-specific VAERS system in the USA (Varricchio et al., 2004). Meanwhile, the Vaccine Ontology (VO) is not suitable for representing complex VAE data. VO has been focused on classification of various vaccines, including licensed vaccines, vaccines in clinical trials, and vaccines only verified in laboratory animal models. VO also represents various types of vaccine components (*e.g.*, vaccine antigens, adjuvants, and vectors), vaccine attributes (*e.g.*, vaccine organism viability and virulence), vaccination methods, and other concise and closely related vaccine information. The inclusion of complex VAE information to VO would make VO imbalanced and lack of focus. Due to these reasons, we generated the VAE-specific OVAE. Since both OAE and VO use the Basic Formal Ontology (BFO) (http://www.ifomis.org/bfo) as the top level class, the alignments between OVAE, OAE, and VO are easy and straightforward.

As an extension of OAE and VO, OVAE targets for not only importing related terms from these two ontologies but also including many OVAE-specific terms. The primary data source for generating vaccine-specific AE ontology terms in current OVAE is the official vaccine package inserts available in the USA FDA website (FDA, 2013). Each official vaccine package insert document provided by the USA FDA includes a section called

"Adverse Reactions". The results provided in the section were obtained from carefully designed clinical trials with randomized controls and worldwide post-marketing experience. Therefore, the VAE information provides basic known VAEs that are likely to occur after an administration of a specific vaccine in a human vaccinee. Based on the officially documented information, OVAE includes many OVAE-specific terms, for example, 'Afluria-associated pain AE' to define a pain AE specific for Afluria-vaccinated patients. As shown in detail later in the paper, the generation of these new terms allows the inclusion of more detailed information about these VAEs, for example, the VAE occurrences in human vaccinee populations in different age groups.

Table 1 lists the OVAE statistics as of May 1, 2013. OVAE used the most recent BFO 2.0 Graz version (http://purl.obolibrary.org/obo/bfo.owl) as the top level ontology. Since BFO 2.0 is not yet finalized, some relation terms (*e.g.*, 'part of' or BFO_0000050) are still used in OVAE but do not necessarily comply with the most recent BFO 2.0. During the process of importing many AE or vaccine-related terms from OAE and VO to OVAE, many terms from other existing ontologies, including OGMS, Ontology for Biomedical Investigation (OBI) (Brinkman et al., 2010), Phenotypic Quality Ontology (PATO) ("PATO - Phenotypic Quality Ontology,"), and Information Artifact Ontology (IAO) (http://code.google.com/p/information-artifact-ontology/), have also been imported to OVAE (Table 1). To maintain the ontology asserted and inferred hierarchies and support intact reasoning capability, the OntoFox software was used for external term importing (Xiang, et al., 2010). In summary, OVAE includes 1,199 terms that contains 652 OVAE specific terms (with "OVAE_" prefix). In addition, OVAE including all 113 terms from the BFO version 2.0, 315 VO terms, 105 OAE terms, 3 OBI terms, 3 IAO terms, and 2 OGMS terms (Table 1). By referencing the vaccine package insert data, OVAE represents 87 distinct AEs associated with 63 licensed human vaccines.

Table 1. Summary of ontology terms in OVAE.

Ontology Names	Classes	Object properties	Data properties	Total
OVAE	650	1	1	652
BFO	35	78	0	113
OBI	2	1	0	3
PATO	7	0	0	7
IAO	3	0	0	3
OAE	105	0	0	105
OGMS	2	0	0	2
VO	307	5	3	315
Total	1110	85	4	1199

3.2 OVAE design pattern of representing VAE

The general design pattern of representing a VAE in OVAE is shown in Fig. 1. Specifically, a licensed vaccine, manufactured by a company and having specific quality (*e.g.*, using inactivated vaccine organism), is targeted to immunize a human vaccinee against infection of a microbial pathogen. A particular vaccination route (*e.g.*, intramuscular route) is specified. A specific VAE (*e.g.*, Afluria-associated injection-site pain adverse event) occurs in a human vaccinee and after (*preceded_by*) a vaccination. The human vaccinee, having a specific age (defined via a datatype) at the time of vaccination, is part of the population of human vaccinees using this vaccine. The VAE occurrence is defined as a frequency of an adverse event associated with the administration of a vaccine in a vaccinee population. The new object property term 'has VAE occurrence' is defined in OVAE to specify a VAE occurrence (xsd:decimal datatype) in a human vaccinee population that has been individually vaccinated with a specific vaccine during a specific time period. To simplify the representation of axioms linking vaccine adverse event and human vaccinee population, OVAE generates a shortcut relation 'occurs in population' (Fig. 1).

The vaccine attributes and vaccination details are imported from VO. Their inclusion in the design pattern is due to their possible contribution to the VAE determination. For example, a live attenuated vaccine and a killed inactivated vaccine may in general induce different types or levels of VAEs, which can be analyzed by statistical analysis (Sarntivijai, et al., 2012).

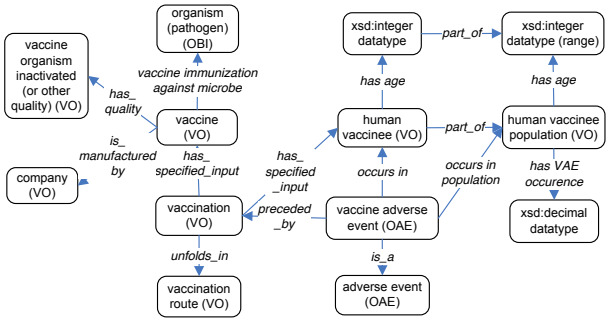

Fig. 1. OVAE design pattern of a human vaccine adverse event.

One novelty in the design pattern is the generation and application of the population term 'human vaccinee population' to define a VAE occurrence. In previous versions of OAE and VO, only 'vaccinee' and 'human vaccinee' (*i.e.*, a human being administered with a vaccine) exist. However, it is incorrect to say that a specific human vaccinee has a VAE occurrence of some percentage (*e.g.*, 10%). An occurrence is defined only for a population. The generation of the term 'human vaccinee population' solves the ontology modeling issue. Any particular human vaccinee is part of a human vaccinee population.

There are two different approaches for representing the relation between a human vaccinee (or human vaccinee population) and an age (or age range). One approach is to link a vaccinee to a quality named 'age', and then link the 'age' to a datatype using the OBI relation term 'quality measured as'. Another approach for representing the relation is to generate a shortcut relation 'has age' (or specifically 'has age in year'). To make the representation simpler and reasoning efficient, we have taken the second choice. An example is provided below (Fig. 2).

3.3 OVAE design pattern of representing VAE

The FDA website includes supporting materials for most human vaccines licensed in the USA (FDA, 2013). To efficiently represent VAEs reported in the package inserts, an MS Excel template was developed with the following categories: vaccine name, vaccine VO ID, VAE location, VAE name in package insert, VAE name in OAE, OAE ID, age category, age years, VAE occurrence, and reference. Data for each category was manually collected from individual vaccine package inserts and then input into the Excel template. The VAE location is listed as either injection-site or systemic. The injection-site location is incorporated as part of the OAE term, while the systemic AEs are set up as default. Age categories included child (typically under 18 years old), adult (above 18 years old), senior (above 65 years old), or child-adult (all ages). Specific ages are concerted to years and presented to comply with the OWL format. Each VAE is referenced by the package insert citation. The data were then imported to OVAE using the Ontorat tool (Xiang, et al., 2012).

An example of OVAE representation of VAE is shown in Fig. 2. Briefly, Afluria has been associated with nine different types of AEs, including injection-site pain AE that has been defined in OAE (Fig. 2A and 2B). For each AE, it is likely that different VAE occurrences are reported based on the age groups. OVAE uses two datatype property terms ('has age in year' and 'has VAE occurrence') to link vaccinee population groups and VAEs associated with particular VAE occurrences (Fig. 2B). The "OR" clause is used to include vaccinee populations with different age ranges. The information matches to the FDA package insert information (Fig. 2C) which is cited as a definition source (annotation property).

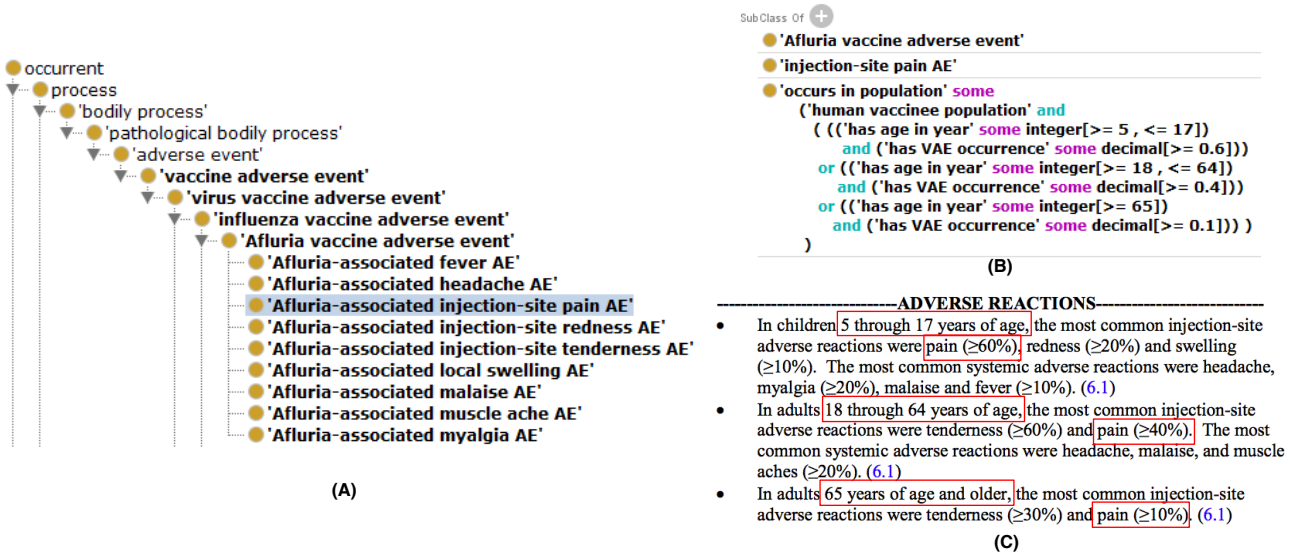

Fig. 2. OVAE representing Afluria VAEs reported in FDA vaccine package insert. (A) The hierarchical structure of Afluria VAEs represented in OVAE. (B) OVAE axiom representation of two types of 'Afluria-associated injection-site pain AE' based on two age groups. (C) Afluria adverse reactions recorded in the FDA package insert document. Other VAEs shown in the FDA package inserts are also represented in OVAE. The subfigures (A) and (B) were screenshots of OVAE using the Protégé OWL editor.

3.4 OVAE VAE data analysis

After all VAEs found in FDA licensed vaccines are represented in OVAE, OVAE was queried using SPARQL. Different questions were addressed via the analysis of the OVAE knowledge base as exampled below.

First, those vaccines that are associated with the largest number of VAEs were analyzed (Table 2). It is interesting that many of these vaccines protect against meningitis, which is caused by different pathogens including *Haemophilus influenza* type b (Comvax and PedvaxHIB) and *Neisseria meningitides* (Menactra). The list also includes two influenza vaccines and two Diphtheria-Tetanus vaccines

(Table 2). It is noted that the information does not dictate the severity of AEs associated with each vaccine, but instead indicates those vaccines that are licensed for human use in the USA and display the most variation in their reported AEs.

Table 2. Top 10 vaccines with the largest variety of VAE reported.

Vaccine Name	VO_ID	Total # VAE
Recombivax HB (Hepatitis B)	VO_0010737	23
Menactra (Meningitis)	VO_0000071	21
Comvax (Meningitis, Hepatitis A)	VO_0000028	19
Prevnar (*Streptococcus pneumonia*)	VO_0000090	19
Tetanus and Diphtheria Toxoids Absorbed by MA Biological (Tetanus, Diphtheria)	VO_0000111	18
Fluarix (Influenza)	VO_0000045	15
Fluarix Quadrivalent (Influenza)	VO_0000983	15
PedvaxHIB (Meningitis)	VO_0000083	15
RabAvert (Rabies)	VO_0000094	14
Boostrix (Tetanus, Diphtheria, Pertussis)	VO_0000015	14

Note: The disease or infection being protected against is specified next to each vaccine name. The vaccines are sorted based on the VAEs recorded in their package insert documents.

Secondly, we evaluated the top VAEs that have been reported most frequently among all vaccines licensed in the USA and represented by OVAE (Table 3). Most of the top 10 frequently observed VAEs are expected, such as injection-site pain and redness, fever, and local swelling. The headache and myalgia (*i.e.*, muscle pain) AEs demonstrate two types of pain. Similar to different pains, malaise (*i.e.*, uneasiness and discomfort) and fatigue are sensory AEs. It is noted that the information does not dictate which VAEs are the most severe, but indicates which VAEs are commonly observed in currently licensed vaccines in the USA.

Table 3. Top 10 most frequently reported VAEs

AE Name	OAE_ID	Total # vaccines	%
Injection-site pain AE	OAE_0000369	43	68.3
Headache AE	OAE_0000377	39	61.9
Fever AE	OAE_0000361	34	54.0
Local swelling AE	OAE_0001139	30	47.6
Injection-site redness AE	OAE_0001546	25	40.7
Irritability AE	OAE_0001105	23	36.5
Malaise AE	OAE_0000390	21	33.3
Injection-site erythema AE	OAE_0000644	20	31.7
Myalgia AE	OAE_0000375	19	30.2
Fatigue AE	OAE_0000034	18	28.6

To better understand the top VAEs associated with licensed human vaccines, the hierarchical structure of the top 10 VAEs (Table 3) was extracted using the tool OntoFox and visualized using Protégé ontology editor (Fig. 3). The hierarchical visualization indicates that most of the top ranked VAEs belong to the behavior and neurological AE branch.

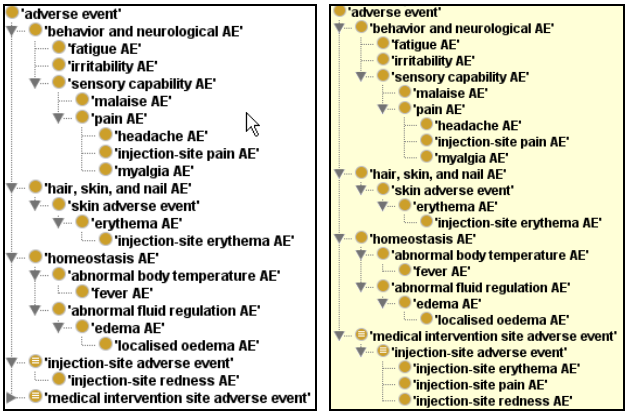

Fig. 3. Classification of top 10 AEs associated with licensed human vaccines in the US. These OAE terms have been imported to OVAE using OntoFox and visualized using Protégé OWL editor. (A) Asserted hierarchy in OAE; (B) Inferred hierarchy after reasoning.

Lastly, we compared the VAEs and VAE occurrences under different age groups. As shown in Fig. 2, the OVAE clearly represents the associations between VAEs, the VAE occurrence rates, and different ages (in years) of human vaccinee population. Our analysis can further identify which age category has a higher probability of experiencing any specific adverse events. For example, we found that *Salmonella typhi* vaccine Typhim Vi is associated with injection-site tenderness adverse events with the highest rate of 97.5% at the age group of 18-40 years old. Based on the classification of "child", "adult", and "child-adult" described in Section 3.3, there are 240, 160, and 177 specific VAEs in the age categories "child", "adult", and "child-adult", respectively. It is also found that in general the VAE occurrences shown in the children are typically higher than those in adults. This suggests that individuals under 18 years may be more likely to experience an adverse reaction after vaccination.

4 DISCUSSION

The development of OVAE is aimed to align and reuse existing ontologies OAE and VO, and systematically represent and analyze vaccine-specific adverse events (VAEs). As demonstrated in this report, such a strategy has many advantages. First, as shown in Fig. 2, the ontological classification is easy for humans to interpret and analyze. A human can browse the hierarchical tree to quickly understand which VAEs are typically associated with a licensed vaccine. Secondly, the ontology OWL representation is also interpretable by computers and software programs. New

programs can be developed to parse and analyze the information. Thirdly, the approach of aligning OVAE with existing ontologies allows efficient integration of data presented in other ontologies (*e.g.*, VO). Such a seamless integration makes it possible to analyze VAEs with other tools such as VO-based literature mining (Ozgur, Xiang, Radev, & He, 2011). In addition to the VAEs associated with USA licensed vaccines, the OVAE can be used to represent VAEs associated with vaccines licensed in other countries.

It is also possible to apply the OVAE framework to analyze clinical VAE data such as those case reports stored in VAERS (Varricchio, et al., 2004). For example, by comparing the reported vaccine-specific VAE cases in VAERS with the VAE occurrences reported in the package inserts and OVAE, it is easy to differentiate known VAEs and possibly new VAEs associated with the vaccine. Many differences exist in terms of the data shown in the package inserts and in VAERS database. While the data in the package inserts were typically obtained from well controlled clinical trials, clinical VAE case reports stored in VAERS came from random reports from physicians, patients, patients' parents, or other sources. The VAERS database does not indicates the total number of vaccinated human vaccinees in any given period, making it impossible to calculate exact VAE occurrences as reported in the package inserts and OVAE. However, ontological approach, together with statistical analysis, is still useful in VAERS data analysis as previously demonstrated (Sarntivijai, et al., 2012). One future research direction will be to identify novel ways to analyze VAE clinical data using OVAE.

While many AEs are common, different vaccines are associated with different AEs with various molecular mechanisms. The ontology representation of vaccine-specific AEs is a first step towards refined deep understanding of vaccine adverse events. It is also noted that the method of establishing vaccine-specific OAE extension may likely be applied for developing OAE extensions in other specified domains such as drug-associated adverse events.

ACKNOWLEDGEMENTS

This project was supported by a NIH-NIAID grant (R01AI081062). We also appreciate the work by Bin Zhao who helped clean up the ontology.

REFERENCES

Brinkman, R. R., Courtot, M., Derom, D., Fostel, J., He, Y., Lord, P., . . . Consortium, T. O. (2010). Modeling biomedical experimental processes with OBI. *Journal of Biomedical Semantics*, June 22; 21 Suppl 21:S27. doi: 10.1186/2041-1480-1181-S1181-S1187.

Cunha, M. P., Dorea, J. G., Marques, R. C., & Leao, R. S. (2013). Vaccine Adverse Events Reported during the First Ten Years (1998-2008) after Introduction in the State of Rondonia, Brazil. *Biomed Res Int, 2013*, 853083. doi: 10.1155/2013/853083

FDA, The United States (U.S.). (2013). U.S. Food and Drug Administration. Vaccines Licensed for Immunization and Distribution in the US with Supporting Documents. URL: http://www.fda.gov/BiologicsBloodVaccines/Vaccines/ApprovedProducts/UCM093830.htm, accessed on April 3, 2013

He, Y., Cowell, L., Diehl, A. D., Mobley, H. L., Peters, B., Ruttenberg, A., . . . Smith, B. (2009, July 24-26). *VO: Vaccine Ontology*. Paper presented at the The 1st International Conference on Biomedical Ontology (ICBO 2009), Buffalo, NY, USA. Nature Precedings: http://precedings.nature.com/documents/3552/version/1.

He, Y., Xiang, Z., Sarntivijai, S., Toldo, L., & Ceusters, W. (2011). AEO: a realism-based biomedical ontology for the representation of adverse events. Adverse Event Representation Workshop, International Conference on Biomedical Ontologies (ICBO), University at Buffalo, NY, July 26-30, 2011. . *Proceeding of ICBO-2011*, 309 - 315.

Lin, Y., & He, Y. (2012). Ontology representation and analysis of vaccine formulation and administration and their effects on vaccine immune responses. *J Biomed Semantics*, *3*(1), 17. doi: 2041-1480-3-17 [pii]10.1186/2041-1480-3-17

Ozgur, A., Xiang, Z., Radev, D. R., & He, Y. (2011). Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. *J Biomed Semantics, 2 Suppl 2*, S8. doi: 10.1186/2041-1480-2-S2-S82041-1480-2-S2-S8 [pii]

PATO - Phenotypic Quality Ontology. from http://obofoundry.org/wiki/index.php/PATO:Main_Page

Sarntivijai, S., Xiang, Z., Shedden, K. A., Markel, H., Omenn, G. S., Athey, B. D., & He, Y. (2012). Ontology-based combinatorial comparative analysis of adverse events associated with killed and live influenza vaccines. *PLoS One, 7*(11), e49941. doi: 10.1371/journal.pone.0049941PONE-D-12-19530 [pii]

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., . . . Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol, 25*(11), 1251-1255. doi: nbt1346 [pii]10.1038/nbt1346

SPARQL query language for RDF. from http://www.w3.org/TR/rdf-sparql-query/

Varricchio, F., Iskander, J., Destefano, F., Ball, R., Pless, R., Braun, M. M., & Chen, R. T. (2004). Understanding vaccine safety information from the Vaccine Adverse Event Reporting System. *Pediatr Infect Dis J, 23*(4), 287-294. doi: 00006454-200404000-00002 [pii]

Xiang, Z., Courtot, M., Brinkman, R. R., Ruttenberg, A., & He, Y. (2010). OntoFox: web-based support for ontology reuse. *BMC Res Notes, 3*, 175. doi: 1756-0500-3-175 [pii]10.1186/1756-0500-3-175

Xiang, Z., Lin, Y., & He, Y. (2012). Ontorat web server for automatic generation and annotations of new ontology terms. . *Proceedings of the International Conference on Biomedical Ontologies (ICBO), University of Graz, Graz, Austria, July 24-27, 2012*., URL: http://ceur-ws.org/Vol-897/poster_812.pdf

Xiang, Z., Mungall, C., Ruttenberg, A., & He, Y. (2011). Ontobee: A Linked Data Server and Browser for Ontology Terms. *Proceedings of the International Conference on Biomedical Ontologies (ICBO), University at Buffalo, NY, July 26-30.* , 279-281.