Proceedings of the Joint Workshop on Semantic Technologies Applied to Biomedical Informatics and Individualized Medicine

(SATBI+SWIM 2012)

Held at the 11th International Semantic Web Conference

(ISWC 2012)

November 12th, Boston, USA

Preface

Two major challenges to the use of digitally encoded biomedical data for improving health are its distributed nature and lack of harmonization [1]. Semantic technologies, including ontologies, terminologies, Uniform Resource Identifiers (URIs), and the Resource Description Framework (RDF), are key to addressing these challenges. By enabling the precise identification of entities and the computable encoding of formal class definitions, semantic technologies enable large-scale semantic normalization of distributed biomedical data sets.

The Joint Workshop on Semantic Technologies Applied to Biomedical Informatics and Individualized Medicine, co-located with the 11th International Semantic Web Conference, brought together researchers, developers, and practitioners who are actively applying semantic technologies and biomedical data to improving health. Five peer-reviewed papers describing original research in this area were presented at the workshop.

- Corrigan, Soler, and Delaney present incremental progress of the "Translational Medicine and Patient Safety in Europe" project, funded by the EU FP7. The focus of the work is a proof of concept infrastructure to support the creation of actionable knowledge within the electronic health record for clinical decision support. The infrastructure is based on evidence from new research findings coupled with contemporary clinical knowledge and practice. The focus of this work is the use of methods from ontology development and statistics to create a consistent model of the evidence-of-association between the clinical and diagnostics cues.
- McCusker et al. present a novel and intriguing architecture called the "Global Health Explorer" for processing Twitter "tweets" to identify the occurrence of terms from biomedical ontologies for the purpose of visual analysis and data exploration. Novel features of the architecture include an approach the authors term "Ontology-as-API", and the integration of a high dimensional data visualization tool called the Data Cube Explorer. This paper highlights how ontologies and terminologies perform a critical role in enabling biomedical Natural Language Processing (NLP) algorithms to richly annotate biomedical and entities and relationships. The approach may compliment other public health data sets such as World Health Organizations' Global Health Observatory (GHO) dataset and the ReDD-Observatory.
- De Waard and Scheider propose the use of an ontology model called ORCA to enable better representation of biomedical argumentations. ORCA is a lightweight ontology to represent observational and interpretational assertions in scientific documents. The paper presents a brief description of the ontology, the motivation behind it, related work, and a few biomedical applications. This paper is highly relevant since the reliability and attribution of biomedical results, data, and information is a critical issue in research. Moreover, the research

highlights an important use of ontologies to model scientific discourse and evidence with the vision of creating computable chains of claims and evidence that explicitly model the consensus, disagreement, and questions necessary for advancing science in a given field.

- Baranya et al. present an approach for improving medical visualization of semantically annotated CT-Images. The approach combines multiple biomedical ontologies and image characteristics to define what is referred to as a Transfer Function (TF). Essentially a TF maps volumetric data into optional properties, and in general, is not easy to define. The proposed framework--ANISE--is a rule-based system that comprises multiple annotators and rules engine to for defining the TFs based on semantic annotations using ontologies such as FMA and RadLex.
- Chniti et al. describe a novel framework for writing business rules using a structured language that should be usable by domain experts while ensuring that the rules involve entities from a formal ontology can be executed over an object-oriented decisions support system. Rules authored in natural language are translated to IRL executable rules. From there, there are BOMs, XOMs, Java objects and WODMs.

Many thanks to all our contributors and participants at SATBI+SWIM 2012 and also the Programme Committee whose feedback has resulted in a fruitful collection of papers, providing added value to current leading edge research.

Also, a special gratitude to Dr. Joanne S. Luciano and Dr. Eric Neumann for accepting our invitation and participate as keynote speakers of SATBI+SWIM 2012.

November 2012

Alejandro Rodríguez-González
Jyotishman Pathak
Mark D. Wilkinson
Nigam H. Shah
Robert Stevens
Richard Boyce
Ángel García-Crespo

References

1. IOM (Institute of Medicine). 2012. *Digital data priorities for continuous learning in health and health care: Workshop Summary.* Washington, DC: The National Academies Press.

Organization

Organizing Committee

Alejandro Rodríguez González, PhD
CBGP-UPM, Spain
Jyotishman Pathak, PhD
Mayo Clinic, USA
Mark Wilkinson, PhD
CBGP-UPM, Spain
Nigam H. Shah, MBBS, PhD
Stanford University, USA
Robert Stevens, PhD
University of Manchester, UK
Richard Boyce, PhD
University of Pittsburgh, USA
Angel García Crespo, PhD
University Carlos III of Madrid, Spain

Program Committee

Helena Deus, PhD
Digital Research Enterprise Institute, Ireland
Jesualdo Tomas Fernandez Breis, PhD
University of Murcia, Spain
Michel Dumontier, PhD
Carleton University, Canada
Mikel Egaña Aranguren, PhD
CBGP-UPM, Spain
Joanne Luciano, PhD
Harvard University, USA
Alan Ruttenberg
University of Buffalo, USA
Oktie Hassanzadeh
IBM Research, USA
M. Scott Marshall, PhD
MAASTRO Clinic, Maastricht
William Hogan
University of Arkansas, USA

Table of Contents

Public Health Surveillance Using Global Health Explorer

James P. McCusker, Jeongmin Lee, Chavon Thomas, and Deborah L. McGuinness

Tetherless World Constellation
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th Street Troy, NY 12180, USA
{mccusj,leej35}@rpi.edu, chavon.thomas@hws.edu, and dlm@cs.rpi.edu
http://tw.rpi.edu

Abstract. We demonstrate an early version of a semantic web tool, Global Health Explorer (GHX), that can be used to conduct public health surveillance using Twitter. Our infrastructure can use any controlled vocabulary to extract term uses in Twitter and supports hypothesis formation and exploration of data sets using visual analysis. The resulting data, gathered in RDF, makes it possible to analyze term usage through both temporal and spatial dimensions. GHX uses the qb.js framework to visualize and explore these data across dimensions, intitially time and location. This allows users of GHX to monitor terms from pre-existing ontologies to conduct public health surveillance. We have prototyped the use of GHX to monitor terms from the NCI Thesaurus related to influenza-like illnesses.

1 Introduction

Analysis of word usage relating to disease outbreaks is a growing field within public health. Early success has been recognized in some visible projects including the Google Flu project. Google has been able to show that certain search terms are highly correlated with reports from the Center for Disease Control's Influenza Like Illness (ILI) Index [1]. However, Google Flu did not publish which terms they have identified, and it is not clear if the terms used are direct or indirect indicators of ILI. Lacking the term set makes it difficult for others to replicate results. We are exploring a more transparent approach and use Twitter as an open platform for mining and identifying indicators that can provide a similar kind of discussion. Monitoring twitter mentions from entire existing vocabularies may allow researchers to identify new terms that can serve as indicators for other diseases as well.

2 Architecture

The Global Health Explorer is composed of two components: qb.js[1] and Skitter[2]. qb.js is a general-purpose visualization and exploration environment for annotated RDF data. It facilitates visualization of multidimensional datasets expressed using the RDF Data Cube Vocabulary (QB). [2] This vocabulary supports descriptions of measures, annotations, and attributes of measured entities. We create what we call a Semantic Data Dictionary (SDD) using the QB vocabulary, types of measures (scalar, ordinal, nominal, or binary)[3], units of measure[4], and provenance [3] about how the measure is used. This SDD is what drives the use of data in a qb (pronounced cube) and allows the qb to understand the datasets that it is given to visualize and navigate. The overall architecture is shown in Figure 1.

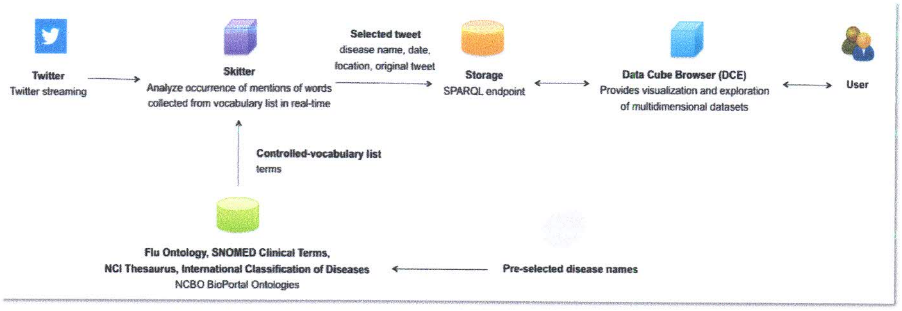

Twitter
Twitter streaming
Skitter
Analyze occurrence of mentions of words collected from vocabulary list in real-time
Selected tweet
disease name, date, location, original tweet
Storage
SPARQL endpoint
Data Cube Browser (DCE)
Provides visualization and exploration of multidimensional datasets
User
Controlled-vocabulary list
terms
Flu Ontology, SNOMED Clinical Terms, NCI Thesaurus, International Classification of Diseases
NCBO BioPortal Ontologies
Pre-selected disease names

Fig. 1. The overall architecture of the Global Health Explorer. Skitter gathers tweets with terms matched from a controlled vocabulary and then stores them in a SPARQL endpoint. qb.js then queries the endpoint to provide a visualization of the data to users

qb.js uses a technique that we are beginning to call ontology-as-API, similar to the use of the term in Kohlhase et al [4]. We use RDFa in an HTML page to encode the configuration of the visualization with minimally sufficient detail to be able to reconstruct it. In Figure 2 we show a starting configuration of qb.js using a particular dataset that is a named graph in a particular endpoint. This architecture allows us to embed useful information, including new SDDs for the data and provenance of visualizations and data that the current visualization has been derived from, into the HTML for the visualization itself. Any further manipulations of the visualization that are then saved and carry the same in-

[1] https://github.com/jimmccusker/qb.js

[2] https://github.com/leej35/Global-Health-Explorer.git

[3] Ontology of Experimental Variables and Values (OoEVV): http://purl.bioontology.org/ontology/OoEVV

[4] Measurement Units Ontology (MUO): http://forge.morfeoproject.org

formation along with it, simply by having it embedded in the RDFa that gets generated on export from qb.js.

(a)

```
<div typeof="http://semanticscience.org/resource/statistical   graph"
     about="#dce" id="dce">
  <span class="config" style="display: none;">
  <a rel="http://www.w3.org/ns/prov#wasDerivedFrom"
      href="http://purl.org/twc/skitter"></a>
  <span typeof="http://rdfs.org/ns/void#Dataset"
        about="http://purl.org/twc/skitter">
    <a rel="http://rdfs.org/ns/void#sparqlEndpoint"
        href="http://localhost:3030/datacube/sparql"></a>
  </span>
  </span>
</div>
```

(b)

```
@prefix prov: <http://www.w3.org/ns/prov#>.
@prefix sio: <http://semanticscience.org/resource/>
<#dce> a sio:statistical   graph;
     prov:wasDerivedFrom <http://purl.org/twc/skitter >.
<http://purl.org/twc/skitter> a void:Dataset;
     void:sparqlEndpoint <http://localhost:3030/datacube/sparql >.
```

Fig. 2. A minimum set of RDFa (a) needed to configure a datacube. DCE uses SIO classes to refer to graphical elements, VOID classes and properties to describe datasets, and PROV-O to show how each uses the other. The id of the DOM node in the HTML defines the element that the DCE visualization will be rendered into. (b) the same abstract RDF graph in Turtle.

Skitter is a tool that searches for mentions of words collected from potentially very large vocabularies. It accesses Twitter to analyze occurrences of terms from a controlled vocabulary in real-time. These concept mentions are then saved into a RDF Data Cube-compatible dataset that is then pushed at regular intervals to any SPARQL endpoint that supports SPARQL UPDATE. The time, user, original text, concepts mentioned, and if available, the location of the tweet is saved in the dataset. Skitter has successfully scaled up to more than 70,000 concepts from the NCI Thesaurus on the sample stream with the only delays occurring on loading and indexing the ontology, which only occurs when the program is started. Figure 3 shows the RDF representation of a tweet that has been extracted by Skitter.

Skitter is able to handle thousands of terms at once by building an in-memory search index of the ontology using Lucene. [5] We use the Snowball stemmer to allow for near matches of terms. Each concept becomes a Document in Lucene with a URI field and as many label fields as are asserted for the concept. As the hash-based index is only built once, on-line searches of terms in the index are constant-time relative to the size of the index. The only limiting factor is the size of the query. Since tweets are limited to 140 characters, this gives a very low upper bound on the time it takes to annotate a tweet. We have been able to successfully process and capture data from the statistical sample Twitter API using a modest laptop.

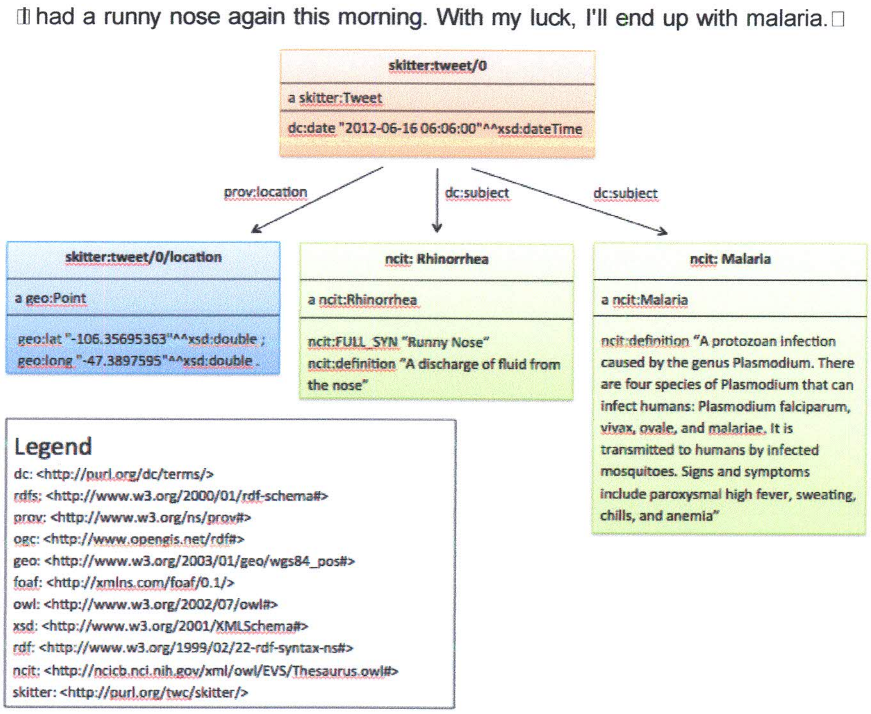

“I had a runny nose again this morning. With my luck, I'll end up with malaria.”
skitter:tweet/0
a skitter:Tweet
dc:date "2012-06-16 06:06:00"^^xsd:dateTime
prov:location
dc:subject
dc:subject
skitter:tweet/0/location
a geo:Point
geo:lat "-106.35695363"^^xsd:double ;
geo:long "-47.3897595"^^xsd:double .
ncit: Rhinorrhea
a ncit:Rhinorrhea
ncit:FULL_SYN "Runny Nose"
ncit:definition "A discharge of fluid from the nose"
ncit: Malaria
a ncit:Malaria
ncit:definition "A protozoan infection caused by the genus Plasmodium. There are four species of Plasmodium that can infect humans: Plasmodium falciparum, vivax, ovale, and malariae. It is transmitted to humans by infected mosquitoes. Signs and symptoms include paroxysmal high fever, sweating, chills, and anemia"
Legend
dc: <http://purl.org/dc/terms/>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prov: <http://www.w3.org/ns/prov#>
ogc: <http://www.opengis.net/rdf#>
geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
foaf: <http://xmlns.com/foaf/0.1/>
owl: <http://www.w3.org/2002/07/owl#>
xsd: <http://www.w3.org/2001/XMLSchema#>
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
skitter: <http://purl.org/twc/skitter/>

Fig. 3. A tweet (skitter:tweet/0) has been identified as having two terms as a subject, Rhinorrhea (runny nose) and Malaria. The tweet has a date (via Dublin Core Terms Date) and a location (prov:location and geosparql).

The Global Health Explorer uses qb.js, skitter, and terms from the NCI Thesaurus [6] and stores them into a single dataset that gradually aggregates in real time. Users can use the qb.js to see how certain terms, at any level, are mentioned in a particular area or time frame.

3 Discussion

The Global Health Explorer allows users to extract any mention of terms from very large vocabularies, such as the NCI Thesaurus, and build detailed datasets. Because many vocabularies are expressed as Linked Data, tools like qb.js can provide additional, on-demand information about the concepts mentioned in tweets simply by dereferencing their URIs. Additionally, use of QB and RDF makes it simple to view datasets in tools like qb.js This is applicable to research areas other than just influenza research, as it can be used whenever discussion of a topic uses more concepts than can fit in a conventional Twitter search. Research into the use of social networks like Twitter for public health surveillance has been very active. The Informatics Research and Development Laboratory (IRDL) recently published research about a web application called PHTweet that examined Twitter streams for a limited number of health terms.[5] Other research has introduced a technique that analyzes large quantities of the queries from the google search engine in order to target the regions of vast populations with influenza epidemics and distinguish the communities with health-seeking habits. [7] In Figure 4 we show a GHX visualization using qb.js.

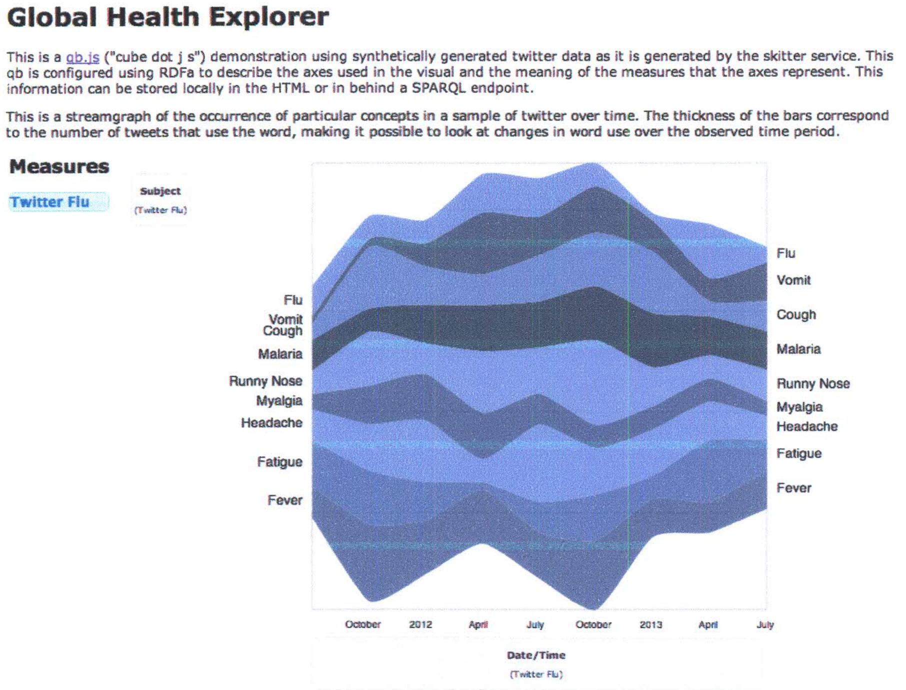
4 Conclusions

We have developed a prototype in which we can demonstrate how users can gather and analyze large data streams for occurrences of particular words (or phrases). These analyses may be useful to spot trends such as the emergence and growth of a disease outbreak. We mine social networks such as Twitter for term mentions and then map those terms into concepts that interpret the meaning of the tweet. We are working to allow researchers and members of the general community who would not ordinarily have access to public health surveillance data to use and access it. Although our Global Health Explorer does not act as a diagnostic system, we hope that it can serve as a new tool for researchers to access and analyze information that would be difficult to gather otherwise.

References

1. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data.” Nature, vol. 457, no. 7232, pp. 1012–4, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19020500
2. R. Cyganiak, D. Reynolds, and J. Tennison, “The RDF Data Cube Vocabulary.” [Online]. Available: http://www.w3.org/TR/vocab-data-cube/
3. T. Lebo, S. Sahoo, and D. McGuinness, “PROV-O: The PROV Ontology.” [Online]. Available: http://www.w3.org/TR/prov-o/

[5] http://edemo.phiresearchlab.org/PHTweet2



Global Health Explorer

This is a qb.js ("cube dot j s") demonstration using synthetically generated twitter data as it is generated by the skitter service. This qb is configured using RDFa to describe the axes used in the visual and the meaning of the measures that the axes represent. This information can be stored locally in the HTML or in behind a SPARQL endpoint.

This is a streamgraph of the occurrence of particular concepts in a sample of twitter over time. The thickness of the bars correspond to the number of tweets that use the word, making it possible to look at changes in word use over the observed time period.

Measures

Fig. 4. This is a streamgraph of the occurrence of particular concepts in a sample of twitter over time. The thickness of the bars correspond to the number of tweets that use the word, making it possible to look at changes in word use over the observed time period. This demonstration is available at http://doppio.med.yale.edu/~jpm78/tw/qb.js/examples/twitter.html. This qb is using synthetically generated twitter data as it is generated by the skitter service. It is configured using RDFa to describe the axes used in the visual and the meaning of the measures that the axes represent. This information can be stored locally in the HTML or in a SPARQL endpoint.

4. M. Kohlhase, J. Corneli, C. David, D. Ginev, C. Jucovschi, A. Kohlhase, C. Lange, B. Matican, S. Mirea, and V. Zholudev, □The Planetary System: Web 3.0 & Active Documents for STEM,□ Procedia Computer Science, vol. 4, pp. 598□607, 2011. [Online]. Available: https://svn.mathweb.org/repos/planetary/doc/epc11/paper.pdf
5. The Apache Software Foundation, □Apache lucene,□ 2006. [Online]. Available: http://lucene.apache.org/
6. G. Fragoso, S. De Coronado, M. Haber, F. Hartel, and L. Wright, □Overview and utilization of the nci thesaurus,□ Comparative and Functional Genomics, vol. 5, no. 8, pp. 648□654, 2004. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447470&tool=pmcentrez&rendertype=abstract
7. G. Eysenbach, □Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet,□ Journal of Medical Internet Research, vol. 11, no. 1, p. e11, 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762766&tool=pmcentrez&rendertype=abstract

Formalising Uncertainty: An Ontology of Reasoning, Certainty and Attribution (ORCA)

Anita de Waard[1] and Jodi Schneider[2]

[1] Elsevier Labs, Jericho, VT A.dewaard@elsevier.com
[2] Digital Enterprise Research Institute, National University of Ireland
jodi.schneider@deri.org

Abstract. To enable better representations of biomedical argumentation over collections of research papers, we propose a model and a lightweight ontology to represent interpersonal, discourse-based, data-driven reasoning. This model is applied to a collection of scientific documents, to show how it can be applied in practice. We present three biomedical applications for this work, and suggest connections with other, existing, ontologies and reasoning tools. Specifically, this model offers a lightweight way to connect nanopublication-like formal representations to scientific papers written in natural language.

Keywords: scholarly communication, ontologies, nanopublications

1 Introduction

Biological understanding is created by scientists collaboratively working on understanding a (part of a) living system. To contribute knowledge to this collective understanding, biologists perform experiments and draw observational and interpretational assertions about these models [19]. In the social and linguistic practice of scientific publishing, the truth value (the confidence in the certainty of a statement), the knowledge source (who stated it) and basis (what was the statement based on) of an assertions are generally indicated through some linguistic expression of certainty or attribution, a 'hedge', of the type 'These results suggest that [A causes B]', or 'Author X implied that [A causes B]', or, in case a proposition is presumed true, the unmodulated '[A causes B]'. In this way, biological papers provide explicit truth evaluations of their own and other authors' propositions and these evaluations and attributions are a core component of shared knowledge constructions.

The goal of the present work is to provide a lightweight, formal model of this knowledge value and attribution, to assist current efforts that offer formal representations of biological knowledge and tie them more directly to the natural language text of scientific papers. Formal knowledge representations (and their corresponding ontologies) generally consist of statements of the system 'A causes B', or 'A is-a B', that do not leave room for doubt, dispute, and disagreement. But if we want to model the process (as opposed to the final consensus of outcome) of science, we need to trace the heritage of claims. Very often, claims can be traced to interpretations of data – so to model claim-evidence networks [6,15] we need to allow for links from claims to non-textual elements such as figures and provenance trails, to trace the attribution of claims to people, organizations, and data processes [8].

Our model is based on an analysis of scientific argumentation from different fields: linguistics, sentiment analysis and genre studies. We have developed a lightweight ontology, dubbed 'ORCA', the Ontology of Reasoning, Certainty and Attribution, for making RDF representations of the certainty and source of claims. The goal of this model is to assist and augment other efforts in bioinformatics, discourse representation and computational linguistics with a lightweight way of representing truth value, basis and source. In this paper, we present our model and show different scenarios for the practical application of this work. We provide a brief overview of related projects, and sketch our thoughts on possible alignments with complementary ongoing efforts.

Following this introduction, in Section 2 we discuss our proposal for representing the strength and source. Then in Section 3 we discuss related work, followed by some realistic application areas in Section 4. We conclude the paper with a discussion of next steps in Section 5.

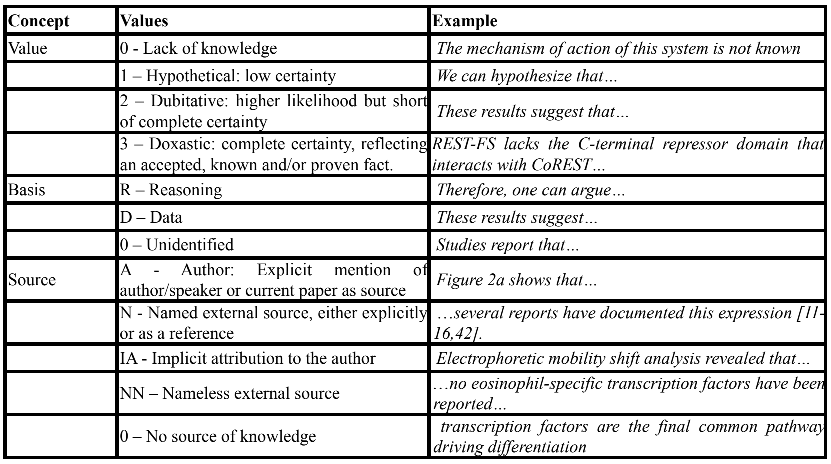
2 Our proposal

2.1 Model

In science, the strength and source of claims are important. Attribution is particularly central in science, yet existing models of provenance do not capture some simple, key distinctions: is the work cited or referred to done by the author or by another person? Is that work backed by data or by inference? Further, the strength of claims is of particular concern, especially during the reviewing process. It is common for authors to need to add qualification to their words, in order to get through the publication process. Even titles must appropriately indicate this in order to get a paper published. For instance, the author proposing the title "miRNA-372 and miRNA-373 Are Implicated As Oncogenes in Testicular Germ Cell Tumors" was instructed to softens the claim by saying that data is the source, making the (un)certainty of this result clearer. To get the paper published, it had to be retitled: "A Genetic Screen Implicates miRNA-372 and miRNA-373 As Oncogenes in Testicular Germ Cell Tumors".

Following concepts in linguistics and computational linguistics (for a full overview of literature, see [7]) we identify a hedged or attributed clause as a Proposition P that is modified by an evaluation E that identifies the truth value and attribution of P. Based on work in linguistics, genre studies and computational linguistics, we identify a three-part taxonomy of epistemic evaluation and knowledge attribution which covers the most commonly occurring types of knowledge attribution and evaluation in scientific text, as shown in the table. This taxonomy is summarized in Figure 1.

From our corpus study [7] it appeared that for all biological statements or 'bio-events' [18] a certain value of E (V, B, S) can be found without much difficulty. Linguistic markers for a lack of full certainty (i.e. where Value ¡ 3) include the use of hedging adverbials and adjectives ('possibly', 'potential' etc), the use of modal auxiliary verbs ('might', 'could') and, most frequently, the use of reporting verbs ('suggest', 'imply', 'hypothesize', etc.). As is clear from the examples in Figure 1, the absence or presence of a single linguistic marker identifies a value in each of the three dimensions:



Concept	Values	Example
Value	0 - Lack of knowledge	*The mechanism of action of this system is not known*
	1 – Hypothetical: low certainty	*We can hypothesize that…*
	2 – Dubitative: higher likelihood but short of complete certainty	*These results suggest that…*
	3 – Doxastic: complete certainty, reflecting an accepted, known and/or proven fact.	*REST-FS lacks the C-terminal repressor domain that interacts with CoREST…*
Basis	R – Reasoning	*Therefore, one can argue…*
	D – Data	*These results suggest…*
	0 – Unidentified	*Studies report that…*
Source	A - Author: Explicit mention of author/speaker or current paper as source	*Figure 2a shows that…*
	N - Named external source, either explicitly or as a reference	*…several reports have documented this expression [11-16,42].*
	IA - Implicit attribution to the author	*Electrophoretic mobility shift analysis revealed that…*
	NN – Nameless external source	*…no eosinophil-specific transcription factors have been reported…*
	0 – No source of knowledge	*transcription factors are the final common pathway driving differentiation*

Fig. 1. Taxonomy

- ‘These results suggest’: Value = 2, Source = Author, Basis = Data;
- ‘REST-FS lacks the C-terminal repressor domain that interacts with CoREST’: Recommended Value = 3, Source = Not specified; Basis = Not specified.

2.2 Ontology

We then model this in a lightweight ontology, ORCA – the Ontology of Reasoning, Certainty and Attribution. From the taxonomy, we have three core aspects: the source of knowledge, its basis, and its certainty. Our ontology should allow us to associate values for each of these. Thus we model them as classes and add associated Object Properties to add flexibility of expression. Controlled values for the taxonomy (e.g. “Named External Source”) are represented as instances. Further, we induce an order on the certainty values, using transitive properties, to make it evident that, e.g. “Hypothetical Knowledge” is less certain than “Dubitative Knowledge”. We considered using SKOS[3] to induce this order; however, skos:broaderThan is not appropriate, and skos Collections add an unwanted layer of complexity.

A clear application of this work is to support and help underpin the relations between formal knowledge representation such as nanopublications, and scientific text. A recent paper in Nature Biogenetics argues that

> Some argue that the rhetoric in articles is difficult to mine and to represent in the machine-readable format. Agreed, but frankly, why should we try? All nanopublications will be linked to their supporting article by its DOI. [17]

We think that adding a layer of epistemic validation with knowledge attribution enables a ‘good enough’ representation of the first level of scientific argumentation: the

[3] `http://www.w3.org/TR/2009/REC-skos-reference-20090818/`

statement and citation of claims. “Frankly, we should try” to do this, since this creates a superior representation of scientific argumentation, and as a bonus, allows us to connect nanopublications at a much more fine-grained level that merely the DOI of the paper that contains the statement. By adding a markup that contains the triple representation of the bioevent, augmented by the ORCA values, an evaluated and traceable network of knowledge can be created within and between documents, that can be represented and reasoned with using the same tools and utlilising the RDF-based standards that are currently being developed for other semantic representation projects such as OpenBEL[4], OpenPhacts [25], Eagle-I[5], and others.

As an example, in another paper on nanopublications, Clare et al (2011) [5] propose that the route to a scientific nanopublication can be facilitated by enabling the annotation of scientific notes or blogs at multiple levels of detail. Specifically, the authors propose that an author annotates a statement such as ‘isoproterenol binds to the Alpha-2 adrenergic receptor’ with a triple, linking the concepts ‘isoproterenol’, ‘Alpha-2 adrenergic receptor’ and ‘binds’ to the CHeBI, UniProt and NCI ontologies, respectively.

Enriching this model, we propose to add an epistemic evaluation to a similar statement: ‘These data demonstrated that [...] isoproterenol modulated the binding characteristics of alpha 2-adrenergic receptors’, which we would represent as follows:

```
@prefix orca: <http://vocab.deri.ie/orca#> .

"isoproterenol modulates binding characteristics
alpha 2-adrenergic receptors"
orca:hasSource orca:AuthorExplicitly ;
orca:hasBasis orca:Data ;
orca:hasConfidenceLevel orca:DoxasticKnowledge .
```

This provides a formal representation of the scientifically relevant aspects–the source of the statement, its basis, and its confidence level; or ORCA could be combined with annotation ontologies. This opens up new possibilities, beyond existing work, as we now discuss.

3 Related Work

There are a wealth of efforts in various fields that aim to represent the argumentation of (biomedical) scientific text, which our work builds on and which we hope this little ontology can support. We will only briefly mention efforts pertaining to scientific discourse efforts and computational linguistics - for a more detailed overview, see [7]).

Semantic Scientific Discourse Regarding semantic scientific discourse work, seminal efforts by the Knowledge Media Institute led by Buckingham Shum aimed to represent scientific sense making and offered ScholOnto, a scientific argumentation ontology, see e.g. [3,13]. In addition to SWAN, Clark et al. and the Annotation Ontology [4] which

[4] http://www.openbel.org/

[5] https://www.eagle-i.net/

aims to capture networks of hypotheses and evidence; this work is currently being combined with work on the Open Annotation framework and experiencing a lively series of developments to enable the creation of a robustly scalable framework for supporting argumentation modeling on the semantic web. Other efforts include SALT [10,11,9], the ABCDE format [1], and ontologies such as CiTO [21] are meant to create environments for authoring and citing specific portions of papers (see also [6] for a summary of these efforts). We believe our work can complement all of these efforts. ORCA can easily be used, alone or in combination with annotation ontologies, in order to link evidence to its source, basis, and confidence level.

Biomedical Informatics Several biomedical informatics systems categorize evidence; for a review, see [2]. We see the modularity as a key advantage: while the Gene Ontology[6] indicates evidence codes for biological inference, these cannot be used without importing the entire ontology. By contrast, domain ontologies could easily incorporate a lightweight, modular RDF ontology such as ORCA. Compared to the Evidence Code Ontology[7], ORCA is more suitable for annotating discourse: for instance, it handles citations to external sources and explicitly indicates confidence levels.

Computational Linguistics Within computational linguistics a number of efforts have focused on detecting the key components and salient knowledge claims in scientific papers, starting with the seminal work of Teufel [22] who developed a system for describing and set of tools to find 'argumentative zones' in scientific papers. Separate efforts to identify epistemically modulated claims and bioevents started with the work of Light et al [14] among (many) others (e.g. [16,23,24]; for a more complete literature overview, see [7]).

4 Possible applications

Improving the evidence behind drug product labels Drug product labels represent drug-drug interactions to help practitioners appropriately prescribe and avoid adverse drug events [2]. Representing the currently known information from the literature is important, yet the certainty of this knowledge varies considerably. Thus, drug-drug interactions are another important use case for ORCA. Information that two drugs interact should be qualified by an indication of what data backs this finding. The level of certainty is indicated in the literature, and representing this would allow different actions to be taken as appropriate. For frail patients, even suspected, unverified drug-drug interactions, could be avoided, as ongoing research confirms the circumstances and certainty of this information. Experimental treatments might accept suspected bad interactions, up to some higher level of certainty.

Data 2 Semantics Use Case As another potential use case, the Data2Semantics project[8] aims to build a semantic infrastructure to connect (and in future, semi-automatically detect and reconstruct) chains linking clinical recommendations to clinical summaries to

[6] http://www.geneontology.org/

[7] http://www.evidenceontology.org/

[8] http://www.data2semantics.org/

the underlying evidence. Still missing from the lightweight ontology being explored by this project is a formalization of the strength and attribution of these clinical recommendations offering another possible use of ORCA. Specifically, we imagine adding an 'ORCA-layer' (either during authoring, or post hoc), to the recommendations provided in clinical trials, so that these can be assessed and directly cited from clinical guidelines. One can then imagine a semantic representation of the clinical finding itself (augmented with an ORCA-structured clause) that can be automatically mined to prepopulate a proposed set of guideline recommendations, that merely need to be checked off my an editor, and can be constantly updated.

Enriching semantic search As a further application of our work is to enrich semantic search platforms: systems that allow search and retrieval subject-object-relationship triples. For instance, MedIE[9] is a triple-based search platform that can be used to find biomedical correlation sentences. But this search does not distinguish between fact and perhaps-fact – nor between novel information and well-known information. We can imagine a number of questions a user might want to answer:

- Give me all completely verified facts about X
- Tell me who found out what about X
- Show me what X is based on?
- Show me all claims which an author says are true, based on their own data. (Such data-based claim knowledge updates have Value = 2 or 3, Basis = Author, and Source = Data; they can be found with using state-of-the-art semantic parsing [20].)

By representing information with ORCA, semantic search engines such as MedIE could provide a better answer to these questions.

5 Next steps for using ORCA

We envision a mixed-initiative approach for applying ORCA to scientific papers. A text mining system (such as mentioned above) would present an author with a tentative list of ORCA-enhanced claims, i.e. a list of the bioevents or key claims in a paper along with a suggested ORCA assignment of the 'veracity value' for each claim. The author (or editor or curator) would then validate both the claim and its 'veracity value', resulting in a set of claims enhanced with ORCA 'veracity values'. Through concept networks such as those proposed in nanopublications or systems such as BEL[10] articles can then be connected by and enriched with these knowledge networks.

Thus, we believe that future collaborations between efforts in semantic web technologies, bioinformatics and computational linguistics can help develop a future where authors can interact with systems that acknowledge and identify their core claims. Similar mixed-initiative systems have already been used to automatically highlight phrases that could be semantically annotated [12]. In particular, we have taken some preliminary steps to identify 'Claimed Knowledge Updates' - a special case of a bioevent that is claimed by the author and based on data - using state-of-the-art semantic parsing [20].

[9] http://www.nactem.ac.uk/medie/

[10] http://openbel.org

One potential roadblock to such a system is that, at least at first, the system output will need to be corrected, which means another step in submission and editing for this already beleaguered author. Another serious issue is that the current system of providing slightly vague, hedged claims serves a social purpose: authors prefer to think that they have made many improvements on the state of the art, and as long as they hedge their statements appropriately, reviewers will let them get away with it. If authors have to make the validity/strength of the claim explicit at the authoring stage, this might introduce a precision in applying truth value that makes all parties uncomfortable. In fact, many reviews mainly concern the degree or strength of the claims made, with the addition of hedging being a frequent demand.

Yet, given the fact that scientific knowledge continues to grow at a dizzying pace, it seems inevitable that sooner or later we will need to represent more exact representations of that knowledge across collections of papers. Widespread use of systems for marking the value, basis, and source of the hedge will help to represent the richness of this knowledge. And there is no particular reason why this model would be limited to the life sciences. As a succinct, simple, and interoperable ontology in that can be used in combination with any RDF-based system, we hope that ORCA can contribute a small building block to what will prove, undoubtedly, to be a collective effort.

Acknowledgements

Thanks to Aidan Hogan for collaboration on the ontology. Jodi Schneider's work was supported by Science Foundation Ireland under Grant No. SFI/09/CE/I1380 Líon2.

References

1. Anita de Waard and Gerard Tel. The ABCDE format enabling semantic conference proceedings. In *ESWC 2006*.
2. R. Boyce, C. Collins, J. Horn, and I. Kalet. Computing with evidence: Part I: A drug-mechanism evidence taxonomy oriented toward confidence assignment. *Journal of Biomedical Informatics*, 42(6):979 – 989, 2009.
3. S. Buckingham Shum, E. Motta, and J. Domingue. ScholOnto: an ontology-based digital library server for research documents and discourse. *International Journal on Digital Libraries*, 3(3):237–248, Oct. 2000.
4. P. Ciccarese, M. Ocana, S. Das, and T. Clark. AO: an open annotation ontology for science on the web. In *BioOntologies 2010*, May 2010.
5. A. Clare, S. Croset, C. Grabmueller, S. Kafkas, M. Liakata, A. Oellrich, and D. Rebholz-Schuhmann. Exploring the generation and integration of publishable scientic facts using the concept of nano-publications. In *Proc. 1st Workshop on Semantic Publishing 2011 at ESWC2011.*
6. A. de Waard, S. Buckingham Shum, A. Carusi, J. Park, M. Samwald, and Á. Sándor Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. Proc. of the Workshop on Semantic Web Applications in Scientific Discourse at ISWC-2009.
7. A. de Waard and H. Pander Maat. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proc. of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55, 2012. Association for Computational Linguistics.

8. P. Groth, Y. Gil, J. Cheney, and S. Miles. Requirements for provenance on the web. *International Journal of Digital Curation*, 7(1):39 – 56, 2012.
9. T. Groza, S. Handschuh, K. Möller, and S. Decker. KonneX-SALT: First Steps Towards a Semantic Claim Federation Infrastructure. In S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, editors, *The Semantic Web: Research and Applications, Proc. of ESWC 2008*, volume 5021 of *LNCS*, pages 80–94. Springer, 2008.
10. T. Groza, S. Handschuh, K. Mller, and S. Decker. SALT - semantically annotated LaTeX for scientific publications. In *The Semantic Web: Research and Applications, Proc. of ESWC 2007*, LNCS. Springer, 2007.
11. T. Groza, K. Möller, S. Handschuh, D. Trif, and S. Decker. SALT: Weaving the claim web. *The Semantic Web: Research and Applications, Proc. of ISWC+ASWC 2007*, 2007.
12. J.L. Fink, P. Fernicola, R. Chandran, S. Parastatidis, A. Wade A, O. Naim, G. B. Quinn, P. E. Bourne. Word add-in for ontology recognition: semantic enrichment of scientific literature. *BMC Bioinformatics*, 24(11):103, 2010.
13. A. D. Liddo, Á. Sándor, and S. B. Shum. Cohere and XIP: human annotation harnessing machine annotation power. In *CSCW (Companion)*, pages 49–50, 2012.
14. M. Light, X. Qiu, and P. Srinivasan. The language of bioscience: Facts, speculations, and statements in between. In *Proc. of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, pages 17–24, 2004.
15. M. S. Marshall, R. Boyce, H. F. Deus, J. Zhao, E. L. Willighagen, M. Samwald, E. Pichler, J. Hajagos, E. Prudhommeaux, and S. Stephens. Emerging practices for mapping and linking life sciences data using RDF a case series. *Web Semantics: Science, Services and Agents on the World Wide Web*, 14:2 – 13, 2012.
16. B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *Association for Computational Linguistics*, volume 45, page 992, 2007.
17. B. Mons, H. van Haagen, C. Chichester, P.-B. t. Hoen, J. T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, and E. Schultes. The value of data. *Nature Genetics*, 43(4):281–283, 2011.
18. R. Nawaz, P. Thompson, and S. Ananiadou. Evaluating a meta-knowledge annotation scheme for bio-events. In *Proc. of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 69–77, 2010. Association for Computational Linguistics.
19. T. Russ, C. Ramakrishnan, E. H. Hovy, M. Bota, and G. Burns. Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case. *BMC Bioinformatics*, 12:351 – 366, 2011.
20. Á. Sándor and A. de Waard. Identifying claimed knowledge updates in biomedical research articles. In *Proc. of the Workshop on Detecting Structure in Scholarly Discourse*, pages 10–17, 2012. Association for Computational Linguistics.
21. D. Shotton. CiTO, the citation typing ontology. *Journal of Biomedical Semantics*, 1(Suppl 1):S6, 2010.
22. S. Teufel. *Argumentative Zoning: Information Extraction from Scientific Articles*. Ph.D., University of Edinburgh, Edinburgh, Scotland.
23. P. Thompson, G. Venturi, J. McNaught, S. Montemagni, and S. Ananiadou. Categorising modality in biomedical texts. In *Proc. of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 27–34, 2008.
24. V. Vincze, G. Szarvas, R. Farkas, G. Móra, and J. Csirik. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(Suppl 11):S9, 2008.
25. A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons. Open Phacts: semantic interoperability for drug discovery. *Drug Discovery Today*, 2012.

Development of an Ontological Model of Evidence for TRANSFoRm Utilizing Transition Project Data

Derek Corrigan[1], Jean-Karl Soler[2], Brendan Delaney[3]

[1]Royal College of Surgeons in Ireland, Dublin, Ireland
derekcorrigan@rcsi.ie
[2]Mediterranean Institute for Primary Care, Attard, Malta
jksoler@synapse.net.mt
[3]Kings College London, London, United Kingdom
brendan.delaney@kcl.ac.uk

Abstract. The development of decision support tools that assist clinicians effectively practice evidence-based-medicine in primary care is dependent on the development of formal models of clinical knowledge. These formal models are a pre-requisite for bridging the knowledge gap that exists between generation of research knowledge and its application in clinical practice. The TRANSFoRm project has developed formal ontological models to represent diagnostic clinical knowledege providing a basis for future development of diagnostic decision support tools. The conceptual validity of the developed models has been tested through representation of diagnostic clinical evidence obtained from literature sources and International Classification of Primary Care Second Edition (ICPC2) coded clinical evidence captured as part of the Transition project. The models provide a basis for future development of decision support tools as part of the on-going TRANSFoRm project. These tools can assist clinicians to formulate and quantify potential diagnoses based on diagnostic cues extracted from patient electronic health records.

Keywords: Ontology, Semantic Web, Evidence-Based-Medicine, Electronic Health Record, Decision Support

1 Introduction

The application of systematic and rigorous approaches to diagnosis through access to the latest available clinical research has long been advocated as one way of contributing to improving patient safety in family practice. The term ‘evidence based medicine’ has been widely associated with such approaches [1]. The effective practice of evidence based medicine implies the existence and use of an up-to-date repository of clinical knowledge. This can be used for interpretation of the diagnostic cues associated with a presenting patient (whether or not this evidence be in electronic format or written) [2]. The challenges in keeping a repository of diagnostic information up to date are similar to the problems of keeping evidence of the effectiveness of treatments up to date. This manifests itself in a delay between the generation of new clinical knowledge from research activities and the timely dissemination of this knowledge into actual clinical practice [3]. Translational medicine advocates the quicker dissemination of research knowledge to clinical practice. It studies the pathways and mechanisms that may optimally provide for the translation of research knowledge into actionable knowledge in clinical practice [4]. One core area highlighted in the study of translational medicine and evidence based medicine has been the need for the development of more formal shared models and coding of clinical data [5]. This can enable quicker dissemination of actionable knowledge via electronic medical record systems. This paper describes how the TRANSFoRm project ('Translational Medicine and Patient Safety in Europe') is working to develop such formal models provided through a dynamically updateable ontology of coded clinical evidence that will support deployment as part of a broader translational medicine platform.

2 The TRANSFoRm Project

The TRANSFoRm project is a five year EU FP7 funded project involving the cooperation of over 20 academic and industry based European research partners. The aim of TRANSFoRm is to develop and evaluate an electronic infrastructure for the 'learning healthcare system' to support both research (epidemiology and clinical trials) and knowledge translation via primary care electronic health record systems [6]. This involves the development of shared models and service infrastructure which allow for the efficient conduct of research. This is coupled with the delivery of decision support tools based on clinical evidence generated from the same electronic sources of primary care data.

2.1 Knowledge Representation in TRANSFoRm

A core element of TRANSFoRm is the development of shared models that allow for representation and exchange of the three distinct types of knowledge: research knowledge, routine healthcare knowledge and actionable knowledge. This relationship is shown in figure 1.

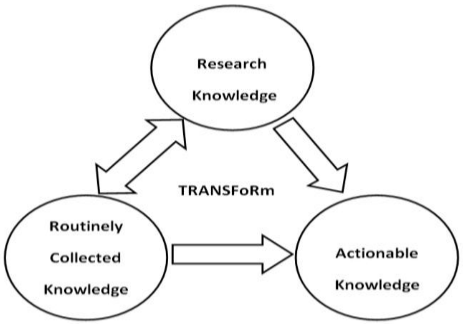


Fig. 1. – Conceptual relationship of Clinical Knowledge types in TRANSFoRm

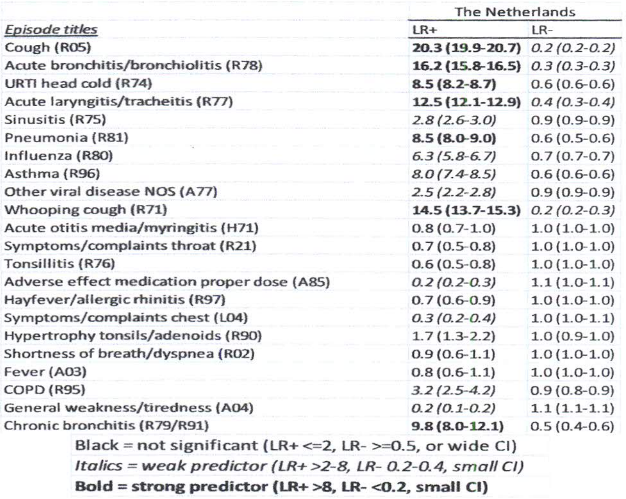
From a decision support perspective the development of a model of actionable clinical knowledge is the core requirement. Actionable knowledge is knowledge that has been distilled from either research knowledge (generated from the conduct of controlled trials and epidemiological studies) and/or from routinely collected healthcare data. It is collected as part of consultations with patients and captured in electronic sources of patient data (such as electronic health records). This requires the application of data-mining and statistical analysis techniques to aggregated sources of electronic patient data to detect trends or patterns in the underlying data that may be used to infer diagnostic association rules. These are then used to construct computable clinical guidelines that can be deployed using decision support tools as part of clinical consultations with patients.

2.2 The Transition Project

The Transition project has demonstrated the feasibility of generating computable actionable knowledge from electronic sources of primary care patient data [7]. A more detailed description of work and methodology used as part of the Transition project has been described elsewhere [8].

The Transition project has utilized the International Classification of Primary Care, second edition (ICPC2) as a clinical classification to provide for the capture of patient data during consultations in family practice in four different countries [9]. The capture of the unambiguous clinical meaning of patient data as recorded in the electronic health record has been recognized as a requirement in the development of formal models of clinical knowledge [10]. A key conclusion of the Transition project was that not only was it feasible to generate actionable knowledge from coded sources of primary care patient data, but that the associations and calculated quantifications of primary care diagnostic cues to diagnostic outcomes were consistent across independent geographic regions.

The Transition project has captured patient data from four countries and quantified the association of ICPC2 coded diagnostic cues to specific diagnostic outcomes. This is presented as calculated likelihood ratios and confidence intervals within the context of a presenting patient reported reason for encounter and geographic region. An example subset of analysis of the association of the symptom ‘cough’ with ICPC2 coded outcomes in the context of the presenting reason for encounter ‘cough’ for patient data collected in the Netherlands is shown in figure 2. The strength of the association is categorized as ‘weak’, ‘strong’ or ‘not significant’ based on the relative value of the likelihood ratio and the width of the associated confidence interval (the methodology and associated calculations are fully described by the Transition Project [8]).



	The Netherlands	
Episode titles	LR+	LR-
Cough (R05)	**20.3 (19.9-20.7)**	*0.2 (0.2-0.2)*
Acute bronchitis/bronchiolitis (R78)	**16.2 (15.8-16.5)**	*0.3 (0.3-0.3)*
URTI head cold (R74)	**8.5 (8.2-8.7)**	0.6 (0.6-0.6)
Acute laryngitis/tracheitis (R77)	**12.5 (12.1-12.9)**	*0.4 (0.3-0.4)*
Sinusitis (R75)	*2.8 (2.6-3.0)*	0.9 (0.9-0.9)
Pneumonia (R81)	**8.5 (8.0-9.0)**	0.6 (0.5-0.6)
Influenza (R80)	*6.3 (5.8-6.7)*	0.7 (0.7-0.7)
Asthma (R96)	*8.0 (7.4-8.5)*	0.6 (0.6-0.6)
Other viral disease NOS (A77)	*2.5 (2.2-2.8)*	0.9 (0.9-0.9)
Whooping cough (R71)	**14.5 (13.7-15.3)**	*0.2 (0.2-0.3)*
Acute otitis media/myringitis (H71)	0.8 (0.7-1.0)	1.0 (1.0-1.0)
Symptoms/complaints throat (R21)	0.7 (0.5-0.8)	1.0 (1.0-1.0)
Tonsillitis (R76)	0.6 (0.5-0.8)	1.0 (1.0-1.0)
Adverse effect medication proper dose (A85)	*0.2 (0.2-0.3)*	1.1 (1.0-1.1)
Hayfever/allergic rhinitis (R97)	0.7 (0.6-0.9)	1.0 (1.0-1.0)
Symptoms/complaints chest (L04)	*0.3 (0.2-0.4)*	1.0 (1.0-1.1)
Hypertrophy tonsils/adenoids (R90)	1.7 (1.3-2.2)	1.0 (0.9-1.0)
Shortness of breath/dyspnea (R02)	0.9 (0.6-1.1)	1.0 (1.0-1.0)
Fever (A03)	0.8 (0.6-1.1)	1.0 (1.0-1.0)
COPD (R95)	*3.2 (2.5-4.2)*	0.9 (0.8-0.9)
General weakness/tiredness (A04)	*0.2 (0.1-0.2)*	1.1 (1.1-1.1)
Chronic bronchitis (R79/R91)	**9.8 (8.0-12.1)**	0.5 (0.4-0.6)

Black = not significant (LR+ <=2, LR- >=0.5, or wide CI)

Italics = weak predictor (LR+ >2-8, LR- 0.2-0.4, small CI)

Bold = strong predictor (LR+ >8, LR- <0.2, small CI)

Fig. 2. – Subset of Transition Project sample analysis for Netherlands showing calculated positive and negative likelihood ratios (LR) with associated confidence intervals (CI)

In the context of the TRANSFoRm project, the output of the analysis generated from the Transition project has provided one useful starting point for informing the modeling of diagnostic clinical evidence as a basis for future diagnostic decision support tool development.

3 Construction of an Ontological Model of Evidence

The usefulness and application of what can broadly be termed 'semantic web' technologies for modeling complex real-world systems has been demonstrated in a wide variety diverse settings including biomedicine, social-networking and on-line retailing [11]. The abstract representation of structures of hierarchical real-world concepts and the definition of the relationships that exist between them is addressed specifically in the area of ontology development. An ontology allows for the development of a portable, shareable, reusable abstract definition of the knowledge domain being modeled [12]. An example of this is the Basic Formal Ontology (BFO) developed as an upper ontology and reused by TRANSFoRm and many other diverse scientific settings [13].

From a clinical perspective, the use of named and bidirectional relationships between ontological clinical concepts enables querying of those concepts from a 'top-down' perspective or a 'bottom-up' perspective. This is useful in modeling of data that would be captured as part of the diagnostic workup process. We can work through our model from a top-level reason-for-encounter down to individual diagnostic cues and back up again in iterative cycles to work through potential differential diagnoses and investigate an individual diagnostic hypothesis [14].

3.1 Ontology Construction Methodology

Many formal methods have been proposed for the design, implementation and validation of ontologies [15-16]. The clinical evidence ontology will be used by a diagnostic decision support application allowing diagnostic workup of potential differential diagnoses to consider based on a presenting patient reason for encounter. A functional approach and definition of a decision support functional specification has driven ontology construction. The expression of functional requirements in the form of informal ontology 'competency questions' was selected as a suitable methodology. This allows formulation of competency questions and their expression using ontology query languages such as SPARQL for testing and validation of defined clinical scenarios.

3.2 Identification of Core Ontological Concepts

A review of the Transition project data identified core ontological concepts that need to be represented in the model. A subset of these showing Transition project definitions for the most important ones and associated examples is shown in table 1.

Table 1. Identified Core Ontological Concepts

General Concept Name	Description and Transition Data Example
Reason for Encounter	An agreed statement of the reason(s) why a person enters the health care system, representing the demand for care by that person. The reason for encounter should be recognized by the patient as an acceptable description of the demand for care. E.g. Cough (as a reason for encounter)
Diagnosis	Formal statement of the providers understanding of the patient's health problem, representing the establishment of an episode of care. It may be a symptom diagnosis or a disease diagnosis. E.g. Chronic Bronchitis
Diagnostic Cue	The symptoms, complaints, objective signs, and/or test results essential for labeling a health problem with a specific diagnosis. E.g. Cough (as a symptom)
Quantification	A quantifiable measure of the association of a diagnostic cue to the presence or absence of a particular diagnosis. E.g. A calculated likelihood ratio value (positive or negative) and associated confidence intervals
Evidence Population	A concept capturing the demographic or population characteristics from which a particular quantification was obtained. E.g. Sex, age, ethnicity or country

3.3 Construction, Population and Hosting Model of Evidence

An ontology of clinical evidence has been constructed for TRANSFoRm using Protégé version 4.1 based on Web Ontology Language (OWL) and Resource Description Framework Schema (RDFS) ontology languages [17-18]. In order to support future development of the decision support tool and to allow for dynamic population of ontology data from analysis done on electronic sources of patient data, the ontology has been deployed to and hosted using the Sesame platform [19]. This provides an open source triple-store backend that has compared favorably in performance tests with other available solutions [20]. It also provides a platform for development and testing of ontology queries using Simple Protocol and RDF Query Language (SPARQL) queries to test the conceptual completeness of the ontology design and the accuracy of generated results. The Transition analysis data for the symptom 'cough' was then manually populated into the ontology.

3.4 Testing and Validating the Model of Evidence

Informal competency questions were translated to formal SPARQL queries to test that all required clinical questions could be expressed using the ontology ensuring conceptual completeness. All generated outputs to those queries were checked for consistency with respect to the original Transition data that was modeled. A sample clinical competency question is: identify the diagnoses for which the symptom X is a strong predictor in the population Y? The formal equivalent query constructed to test for the symptom instance 'cough' in the context of the population instance 'Netherlands' and the associated test result is shown in table 2.

Table 2. Sample SPARQL Formal Query and Results

Formal SPARQL Query	Result (Concept Instances)

`SELECT ?anyDiagnosis`
`WHERE {Cough hasQuantification ?anyQuantification.`
`?anyQuantification hasPosLREvidenceStrength "Strong predictor"^^xsd:string.`
`?anyQuantification hasEvidenceCountry Netherlands.`
`?anyQuantification hasQuantificationDiagnosis ?anyDiagnosis.}`
	`Cough`
	`AcuteBronchitis`
	`URTIHeadCold`
	`AcuteLaryngitis`
	`Pneumonia`
	`WhoopingCough`
	`ChronicBronchitis`

The query result is correct with respect to the original Transition project data shown previously in figure 2. The characteristics associated with these results could be investigated further using additional SPARQL queries based on the ontology concepts and relationships. The complete list of clinical competency questions developed was successfully translated into equivalent formal SPARQL queries and tested against the host platform to ensure conceptual validity and accuracy of results against the original Transition project data.

4 Future Work

The work done to date has focused on development of a back-end model of evidence and a hosting platform. Initial work is now starting on building a web based clinical evidence service application around this. The web service will support two major interfaces: a query interface for asking diagnostic clinical questions to the web service, and an update interface to allow for regular update of the ontology evidence as generated from data mining and analysis modules applied to aggregated sources of primary care data such as the Transition project. The final stage of work will involve the development of the actual decision support tool. This tool will be integrated with a primary care EHR system to be triggered based on the reason for encounter to collect ontologically controlled diagnostic cues.

5 Conclusions

The ontology models of general evidence developed as part of TRANSFoRm were conceptually descriptive enough to model the ICPC2 based data analysis of the diagnostic associations with the symptom ‘cough’ in the context of four separate population regions. By carrying out additional data mining and analysis on more diagnostic cues it is feasible to develop a full picture of ICPC2 coded diagnostic cues and their associations that have been also been quantified using likelihood ratios based on the underlying patient data that is population specific. The Sesame platform provides a suitable ontology hosting mechanism that TRANSFoRm will utilize to develop a back end web based evidence service to provide decision support. This will be based on evidence generated from electronic sources of primary care data that will be populated or changed dynamically as that underlying patient data grows or changes. This is consistent with the goal of implementing translational and evidence based decision support based on the electronic health record.

References

1. Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B., Richardson, W.S.: Evidence based medicine: what it is and what it isn’t. BMJ. 312, 71–72 (1996).
2. Greenhalgh, T.: How to Read a Paper: The Basics of Evidence-Based Medicine. John Wiley & Sons (2010).

3. Haines, A., Jones, R.: Implementing findings of research. BMJ. 308, 1488–1492 (1994).
4. Grol, R., Grimshaw, J.: From best evidence to best practice: effective implementation of change in patients' care. Lancet. 362, 1225–1230 (2003).
5. Lang, E., Wyer, P., Haynes, B.: Knowledge Translation: Closing the Evidence-to-Practice Gap. Annals of Emergency Medicine. 49, 355–363 (2007).
6. TRANSFoRm, http://www.transformproject.eu
7. The Transition Project, http://www.transitieproject.nl
8. Soler, J.K., Okkes, I., Oskam, S., van Boven, K., Zivotic, P., Jevtic, M., Dobbs, F., Lamberts, H., for the Transition Project: The interpretation of the reasons for encounter 'cough' and 'sadness' in four international family medicine populations. Informatics in Primary Care. In press (2012).
9. WHO International Classification of Primary Care, Second edition (ICPC-2), http://www.who.int/classifications/icd/adaptations/icpc2/en/index.html.
10. Coiera, E.: Guide to Health Informatics. Hodder Arnold Publishers (2003).
11. Kashyap, V., Bussler, C., Moran, M.: The semantic web: semantics for data and services on the web. Springer-Verlag New York Inc (2008).
12. Allemang, D., Hendler, J.: Semantic Web for the Working Ontologist, Second Edition: Effective Modeling in RDFS and OWL. Morgan Kaufmann (2011).
13. Basic Formal Ontology homepage, http://www.ifomis.org/bfo/home
14. Kostopoulou, O.: Diagnostic Errors: Psychological Theories and Research Implications. In: Arts, B.H. of M. and the and Development, A.S. of P.C.R. and (eds.) Health Care Errors and Patient Safety. pp. 95–111. Wiley-Blackwell (2009).
15. Grüninger, M., Fox, M.S.: Methodology for the Design and Evaluation of Ontologies. In: Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal (1995).
16. Fernandez-Lopez, M., Gomez-Perez, A.: Overview and analysis of methodologies for building ontologies. The Knowledge Engineering Review. 17, 129–156 (2002).
17. The Protégé Ontology Editor and Knowledge Acquisition System, http://protege.stanford.edu
18. OWL 2 Web Ontology Language Document Overview, http://www.w3.org/TR/2009/REC-owl2-overview-20091027
19. openRDF Sesame homepage, http://www.openrdf.org
20. Bioontology.Org, http://www.bioontology.org/wiki/images/6/6a/Triple_Stores.pdf

A Workflow for Improving Medical Visualization of Semantically Annotated CT-Images

Alexander Baranya[1,2], Luis Landaeta[1,2], Alexandra La Cruz[1], and
Maria-Esther Vidal[2]

[1] Biophysic and Bioengeneering Applyed Group
[2] Semantic Web Group
Simón Bolívar University, Caracas, VENEZUELA
{abaranya,llandaeta,alacruz,mvidal}@ldc.usb.ve

Abstract. RadLex and Foundational Model of Anatomy (FMA) ontologies represent anatomic and image characteristics, and they are commonly used to annotate and describe contents of medical images independently of the image acquisition method (e.g., CT, MR, or US). We present ANISE, a framework that implements workflows to combine these ontologies and image characteristics into Transfer Functions (TFs) that map volume density values into optical properties. Semantics encoded in the image annotations is exploited by reasoning processes to improve accuracy of TFs and the quality of the resulting image.

1 Introduction

In the Life and Health Sciences domains large ontologies have been defined, e.g., SNOMED[3], MesH[4], RadLex[5], and Foundational Model of Anatomy (FMA) [9]. These ontologies are commonly applied to encode scientific knowledge through annotations of concepts, e.g., MeSH terms have been used by curators to annotate and describe PubMed[6] publications and clinical trials published at the Clinical Trials website[7]. Knowledge encoded in these annotations as well as the properties derived from reasoning tasks are used to recovery or discovery properties of the annotated concepts. In this paper we propose a workflow to annotate medical images with terms from RadLex and FMA, and illustrate the benefits of exploiting these annotations during image visualization. We aim at enriching transfer functions (TFs) with semantics encoded in these annotations and provide more precise renderings of the volumetric data of a medical image.

A transfer function (TF) maps density values of volumetric data or voxel into optical properties (e.g., opacity and color) used by rendering algorithms to produce a final image. TFs allow to pre-classify different tissues in an image, and

[3] http://www.nlm.nih.gov/research/umls/Snomed/snomed_mail.html
[4] http://www.nlm.nih.gov/mesh
[5] http://www.rsna.org/radlex/
[6] http://www.ncbi.nlm.nih.gov/pubmed
[7] http://clinicaltrials.gov/

they are based on existing characterizations of the organs that relate a medical image acquisition modality, a tissue, and a density range [7]. Nevertheless, some tissues belonging to different organs may have overlapped densities, and specifying a TF will normally require a robust segmentation technique and specialized segmentation processes to produce a precise tissue classification able to distinguish tissues with overlapped densities. Recently, the problem of tissue classification by semantically annotating volumetric data has gained attention in the literature [2, 3, 5, 8]. Rautek et al. [8] present a fuzzy rule-based system that maps volumetric attributes to visual styles; rules are defined by users without representing special knowledge about the rendering technique. Gerl et al. [5] overcomes this limitation and propose a rule-based system for semantic shader augmentation; this system automatically adds rule-based rendering functionality to static visualization mappings in a shader program. Although both systems rely on rule-based systems to characterize TFs, they do not exploit knowledge encoded in ontologies to improve the quality of the visualization process. Möller et al. [6] present a technique for annotating and searching medical images using ontological semantic concepts for retrieving images from a Picture Archiving and Communication System (PACS); ontologies as FMA and RadLex are used to retrieve data, however, they are not exploited during visualization or tissue classification from the image data. Although applications of semantic annotations have been illustrated, nothing is said about the benefits of using these annotations and the encoded semantics during the definition of TFs.

We present ANISE (an ANatomIc SEmantic annotator), a framework for specifying TFs based on semantic annotations. TFs are based on pre-elaborated semantic annotations of volumetric data which are validated against existing medical ontologies. ANISE relies on a customized reasoner to infer the bounding boxes which contain organs or tissues of a given sub-volume area, as well as its main properties, e.g., density and opacity. Knowledge encoded in the ontologies contribute to characterize and locating tissues by applying specific organ selection algorithms; thus, voxels that are not part of the organ of interest are not considered during the classification process.

This paper contains four additional sections. Section 2 describes ANISE and Section 3 illustrates the ANISE workflow. Section 4 discusses the observed results, and we conclude in Section 5 with an outlook to future work.

2 Architecture

Achieving high quality image rendering requires interpreting each intensity value according to a given tissue. In consequence, a correct representation of information through semantic annotations should ensure: i) minimal error tissue classification due to reasoning and inference, and ii) an accurate visual representation. Figure 1 shows the main components of ANISE: an annotator, a rule-based system, and a visualization module. The Annotator extends an image original annotations with terms that encode the properties of the classified tissues. The rule-based system relies on inference tasks to process

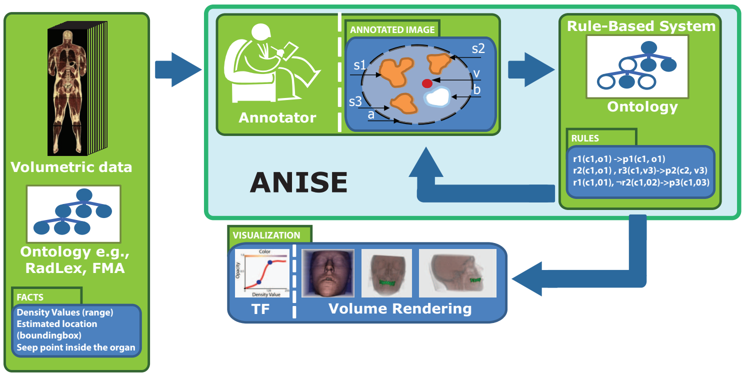


Fig. 1. The ANISE architecture.

original annotations and derive facts that will be used to annotate an image. Annotations regarding to visualization methods and anatomic parts are inferred using Ontology relations (e.g., Subclass) for specific classes (e.g., the Anatomical Set). Finally, the Visualization module executes visualization algorithms on the annotated volumetric data.

Annotator: annotates an image with information about: i) resource authoring, type and identification; ii) acquisition modality; iii) acquisition characteristics like patient orientation in the image; iv) structural and anatomic elements presented and identified in the image; v) regions and points of particular interest; and vi) rendering information. ANISE relies on the following ontologies to extend original image annotations:

- ***Foundational Model of Anatomy*:** FMA allows to describe membership and spatial relationships among voxels in the volume to infer new facts. Furthermore, there are terms in this ontology that can be used for annotating non-anatomical elements, e.g., bounding boxes around particular anatomical organs or some particular points of interest.
- ***RadLex*:** RadLex is an ontology defined for radiologist; it is composed of terms required to annotate medical images. ANISE relies on RadLex terms to describe characteristic from the image itself such as modality, and other acquisition related characteristics that may alter the interpretation and visualization of an image, e.g., orientation.

Rule-Based System: annotations are used during the inference process to derive new annotations. First, it analyses the image acquisition characteristics and correlates body structures of particular interest in order to normalize information for further processing. A bounding box method is used to model anatomic information [3]. Then, combining this information with tissue pre-classification, the inference process is expressed in Probabilistic Soft Logic (PSL) [1]; this process determines the likelihood for a given tissue to be included in a particular region. Closely located tissues with similar intensity values are usually treated as the same values; thus, spatial and anatomic information is used to discriminate by annotating specific points; segmentation based on

voxels neighborhood represent these tissues considering the associated semantic annotations. Ontology classification reasoning tasks are performed with Jena[8].
Visualization Module: derived annotations are used by rendering algorithms to visualize the classified tissues. Partial piece-wise transfer functions are used to select appropriate color and opacity values and rendering them. Default transfer functions are only applied on non-annotated voxels and regions.

3 Applying an ANISE Workflow- A Use Case

We illustrate the ANISE workflow in three different datasets (Table 1), to visualize the FMA term *dentition* from a CT-Head volume data.

Volume Data	Dimensions (voxels)	Voxel size (mm)	File size (MB)
skewed_head.dat	184x256x170	1x1x1	16.0
visible_head.dat	512x512x245	1x1x1	128.0
ct_head.dat	256x256x113	1x1x2	14.8

Table 1. Datasets used for illustrating the utility of using semantic annotations on Medical Images. These datasets are available in [10], [11] and [4] respectively.

Figure 2(a),(d),(g) illustrate the rendering of the images applying a simple TF that maps density values to visualize the tissues that have the same density that dentition; these tissues are colored in *green*. Although data were properly pre-classified, it is not possible to discriminate only dentition by just considering the corresponding densities, i.e., some other tissue were painted, and it was not possible further tuning the TF. In this case the density value range for identifying the dentition overlaps with density value range of other tissues like bone for example. Nevertheless, if semantic annotations are used in conjunction with knowledge encoded in the FMA and RadLex ontologies, ANISE can determine that only the teeth should be colored different than the rest (*green* in our example); this is done by selecting appropriate set of points, applying **Normalization** rules, and considering the **Image Modality** taxonomy. Thus, a better classification for different tissues can be done in an automatic way.

- **Image Modality:** supports a generic tissue classification process which is independent on the image modality. The RadLex term used for *Tomography* is `RID28840`[9] and the term `RID10311` (imaging modality) can be reached by using the SubClass relationship. Further, whenever the image is an MRI the term `RID10312` from the same taxonomy is used to annotate the image, i.e., terms `RID28840` and `RID10312` share an ancestor `RID10311`. Tissues' density ranges are represented as facts and used during the inference process in conjunction with these annotations to pre-classify the image voxels.
- **Volume format:** ANISE current version receives images in raw format, i.e., data correspond to a sequence of intensity values. This information is recovered from the attribute `format` from DCMI[10] metadata.

[8] http://jena.apache.org/

[9] http://purl.bioontology.org/ontology/RID/RID28840

[10] http://dublincore.org/documents/dcmi-terms/

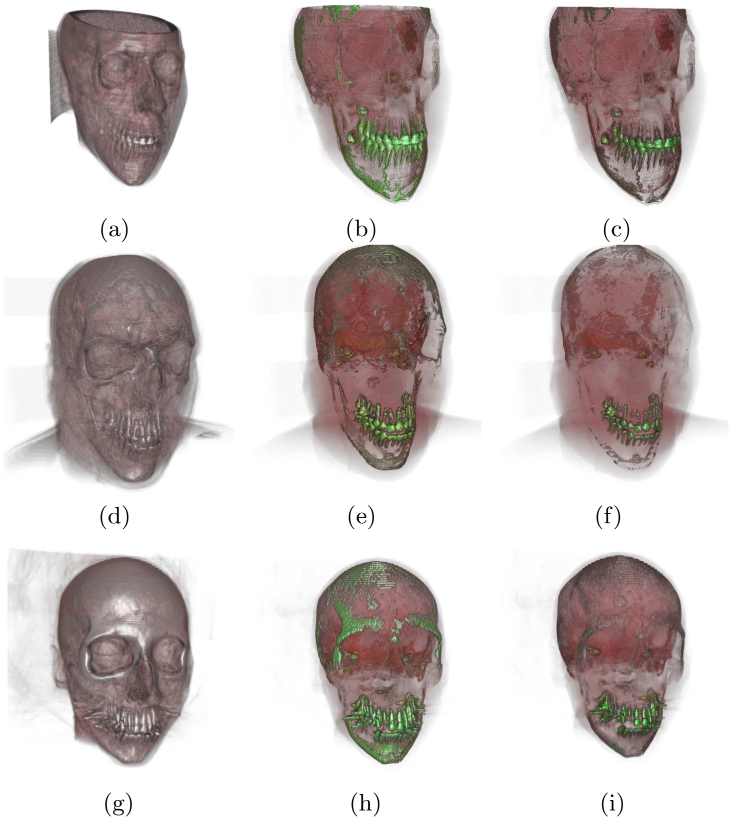


Fig. 2. Results of running the proposed approach with three different datasets: (a) skewed_head.dat, (d) visible_head.dat and (g) ct_head.dat. Images (b), (e), (h) results from rendering without annotation and using a simple TF. Images (c), (f), (i) results from rendering with the application of rule (1) and a semantically enhanced TF.

- **Normalization** rules: are used to transform volumes into a uniform scale considering orientation, voxel size, and modality. Default values are assumed if they are not given. In our use case, we used the term *voxel geometry* `RID2903` from RadLex and its ancestors in the *subClass* branch, i.e., *non-isotropic voxels*, *near-isotropic voxels*, *isotropic voxels.*
- **Dimension:** we used the term *location* (`RID39038`) from RadLex to represent header size, and dimensions in x, y and z of the volume.
- **Tissue:** *dentition* from FMA is the most relevant term in our use case.

We chose *dentition* because it is characterized as the tissue with the higher density value, and the challenge consists on separating the dentition tissue from tissues around it. PSL rules are used to compute the degrees of membership of a voxel to the tissue of interest (*dentition*); it is mainly based on the density value range. The rules that comprise the rule-based system are as follows; they specify

TFs that better visualize the tissue of interest:

$$tissue(X, Y, Z, I) \wedge inside(X, Y, Z, R) \wedge inOrgan(X, Y, Z, I) \rightarrow opacity(X, Y, Z). \tag{1}$$

where, truth values of $opacity(X, Y, Z)$ are determined by the sum of truth values of the following predicates:

- $tissue(X, Y, Z, I)$ describes truth values of the voxel X, Y, Z with intensity I that belong to the objective tissue. This value is defined by:

  $$baseVoxel(X, Y, Z, I) \wedge tissueMap(D, I) \rightarrow tissue(X, Y, Z, I). \tag{2}$$

  where, *baseVoxel(X,Y,Z,I)* is a fact; *tissueMap(D,I)* is a PSL predicate that assigns to an objective tissue D (e.g., *dentition*) the probability of the voxel *X,Y,Z* belongs to the density value range. Initially a density value range is specified, and as far as the inference over the annotations are generated, a new density value range is produced and then, a more precise TF is defined.
- *inside(X,Y,Z,R)* describes truth values of the voxel X, Y, Z belonging to a region R. Applying the inference process, a bounding box that best fits the area of the tissue of interest is derived from an initial location.
- *inOrgan(X,Y,Z,I)* describes truth values of the voxel X, Y, Z belonging to the same organ with intensity I. This value is defined by the rule:

  $$baseVoxel(X, Y, Z, I) \wedge seed(X, Y, Z) \rightarrow inOrgan(X, Y, Z, I). \tag{3}$$

  Given a seed point (*seed(X,Y,Z)*), known to be part of the tissue of interest and analyzing its neighborhood, the area around this seed point is augmented. A point will be part of the tissue if its density value is inside the density value range of the tissue, and close to the tissue area.

Finally, some facts that need to be defined for each dataset are the following:

- **Density value range:** a density value range can be specified initially; however, it can be adapted according to results inferred from the rules.
- **Seed point:** this is a fix value, received from the user describing a voxel known to be part of the tissue of interest.
- **Bounding box:** the rule-based system identifies from an input bounding box, one that better fits the tissue of interest.

4 Discussion

ANISE just considers the most likely localization of a given tissue. First, an initial and basic TF is defined for a normalized model. Then, this model is used for further inferences. Thus, rules are applied independently to the acquisition method by selecting when a density value for a given point in the space falls inside an appropriate interval. As previously stated, simple density classification is not enough to properly determinate matching between voxels of a same tissue or anatomical organ. Additional inference processes need to be conducted; they depend on the annotations. In this example, the region of interest that describes the tissue to be analyzed is presented. A first approach consists of selecting the most likely location of a region of interest, i.e., a bounding box covering the organ of interest. Also, PSL predicates are considered as a possible better approximation of this region with non-zero probability. This is done by considering the neighborhood around the region of interest and knowing that *dentition*, for example, should not be located around eyes or upper areas of the head; voxels belonging to *dentition* should be closer around an area, and distance between dentition voxels should not be longer than certain threshold. Another inference process to adjust the probability for points is performed by considering knowledge derived from ontology relationships, i.e., the classification of the term *dentition* in the Anatomical Set branch. Considering the subClass transitive property (see Figure 3), a seed point is annotated to identify a set element. Then, the voxel neighborhood detection algorithm is performed using PSL predicates. Finally combining all inferred facts and probabilities for given points, likelihood of points that represent a particular tissue are estimated; Figure 4 illustrates the whole process. Further, appropriate TFs for each region are defined and performed. This is done just using the same TF (Fig. 2(b),(e),(h)) but performing a reasoning task that allows to detect the voxels that semantically do not correspond to the tooth tissue and that should not be included in the final volume rendering (see Fig. 2(c),(f),(i)).

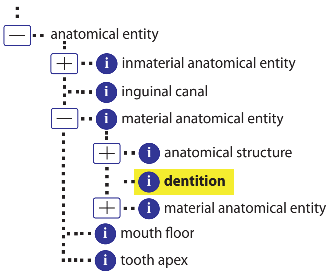


Fig. 3. Scheme from FMA ontology, identifying the Class and SubClass for dentition.

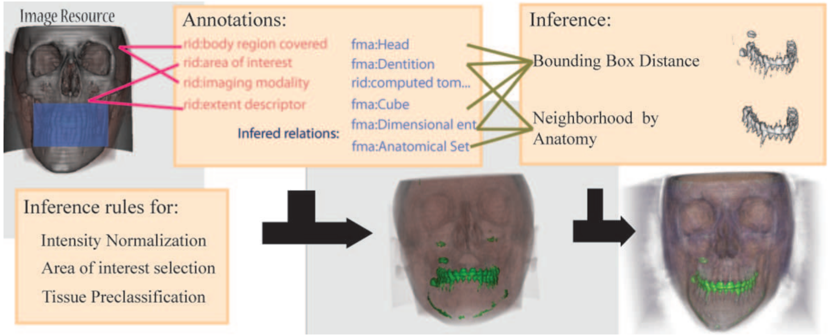


Fig. 4. The ANISE Workflow

5 Conclusions and Future Work

We present ANISE, a framework that exploits knowledge encoded by annotations of 3D medical images, and enhances the rendering process of the images. Quality of ANISE renderings have been studied in different images, and we have observed that they can accurately locate tissues that comprised a medical image. Annotations allow identifying or validating patterns on images, accurate image retrieval, and applying the visualization process on regions of interest. Methods to filter relevant information have been developed at high abstraction level, allowing extension of the inference process to perform particular algorithms, i.e., voxel neighborhood predicates could be improved to allow different methods. In the future, we plan to enhance the rule-based system to normalize a wider range of conditions, and include different image modalities (e.g., MR, and PET) as well as tissues (e.g., blood vessels). Furthermore, we will extend tissue identification algorithms and rules to: *i*) detect and annotate anomalies, and *ii*) identify special conditions on tissues inside the region of interest. Development of visualization algorithms to consider not only TF definitions but also different interpretations of semantic annotations of particular tissues of interest and its corresponding representation on rendered image is also part of our future work.

References

1. M. Broecheler, L. Mihalkova, and L. Getoor. Probabilistic similarity logic. In *Conference on Uncertainty in Artificial Intelligence*, 2010.
2. A. Criminisi, J. Shotton, and S. Bucciarelli. Decision forests with long-range spatial context for organ localization in ct volumes. In *MICCAI workshop on Probabilistic Models for Medical Image Analysis (MICCAI-PMMIA. Springer)*, 2009.
3. A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7(2-3), 2012.
4. http://www-graphics.stanford.edu/data/voldata/CThead.tar.gz.
5. M. Gerl, P. Rautek, T. Isenberg, and E. Gröller. Semantics by analogy for illustrative volume visualization. *Computers & Graphics*, 36(3):201–213, 2012.
6. M. M´oller and S. Mukherjee. Context-driven ontological annotations in dicom images: Towards semantic pacs. In *Proceedings of International Joint Conference on Biomedical Engineering Systems and Technologies*, 2008.
7. B. Preim and D. Bartz. *Visualization in Medicine: Theory, Algorithms, and Applications.* The Morgan Kaufmann Series in Computer Graphics., 2007.
8. P. Rautek, S. Bruckner, and E. Gröller. Semantic layers for illustrative volume rendering. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1336–1343, 2007.
9. C. Rosse and J. Mejino. The foundational model of anatomy ontology. In *Anatomy Ontologies for Bioinformatics: Principles and Practice.* The Morgan Kaufmann Series in Computer Graphics., 2007.
10. http://www.cg.tuwien.ac.at/courses/Visualisierung/1999-2000/skewed_head.zip.
11. http://mri.radiology.uiowa.edu/VHDicom/VHMCT1mm/VHMCT1mm_Head.tar.gz.

Pharmaceutical Validation of Medication Orders Using an OWL Ontology and Business Rules

Amina Chniti[1,2], Abdelali BOUSSADI[2,4,5,6], Patrice DEGOULET[2,4], Patrick Albert[1], Jean Charlet[2,3]

[1] CAS France, IBM
{amina.chniti,albertpa}@fr.ibm.com
[2] Inserm UMRS 872 q.20, Ingénierie des Connaissances en Santé, Paris, France
Jean.Charlet@upmc.fr
[3] AP-HP, Paris, France
[4] Hôpital Européen George Pompidou, Paris, France
{abdelali.boussadi,patrice.degoulet}@egp.aphp.fr
[5] UPMC Université Paris 06, Paris, France
[6] Université René Descartes Paris 05, Paris, France

Abstract. Ontologies are description of domains encoded in a formal language while Business Rules are description of business policies encoded in a natural controlled language. In this paper we present an application of pharmaceutical validation of medication order based on an OWL ontology and business rules or more specifically clinical decision rules. This application has been developed based on a prototype that enables business users to author, in a controlled natural language, execute and manage their Business Rules over OWL Ontologies.

Keywords: OWL Ontology, Business Rule, Clinical Decision Rule, Pharmaceutical Validation.

1 Introduction

Ontologies are more and more used to model the business knowledge which is due to their power of expressiveness and to their flexibility. On the other hand, many business applications are nowadays built based on Business Rules especially after the emergence of the BRMS (Business Rules Management System). Business Rules are a description of a business policy, encoded in a natural controlled language. They define or specify constraints of some aspect of the business[7] and enable automating business decisions. An example of a business rule is given in the following :

IF the presentation name of the drug is "GLUCOPHAGE 850MG TAB"
and the dosage unit of the dosage regimen phase is "TABLET"
THEN the prescripion is not valid;

In this paper, we present an application of pharmaceutical validation that enable to automate the decision of validation of medication orders. It is based on

[7] http://www.businessrulesgroup.org

an OWL ontology and business rules, or more specifically clinical decision rules. The OWL [8] (Web Ontology Language) ontology models the most pertinent entities (concepts and properties) of pharmaceutical validation activity used in the Hôpital Européen Georges-Pompidou (HEGP) (Georges Pompidou European Hospital) [9]. The rules, test on the values given to the entities described in the ontology and assert if a medication order is valid or not [2].

The fact of using business rules enables the business user (*i.e.*pharmacist, physician) to be involved in the implementation of the application as he/she can author the rules in a natural controlled language.

Business rules and ontology have already been combined to support clinical decision [6] [7]. However, end user involvement in the design and the implementation of the application is a neglected aspect. In this study we propose to involve the end user (pharmacists, physicians, nurses) in the implementation of the application and to experiment the business rules designed as a clinical decision rules.

To develop this application, we first implement a prototype [10], *OWL plug-in for WODM*, that enable authoring and executing business rules over OWL ontologies [3]. For this, we based on the infrastructure offered by the Business Rule Management System (BRMS) WebSphere Operational Decision Management (WODM) [11] and added as input OWL ontologies.

This paper is organized as follows; Section 2 present the *OWL plug-in for WODM.* Then Section 3 describes the application of pharmaceutical validation of medication order. Finally Section 4 concludes and presents our perspectives.

2 Proposed Approach

WODM offers an infrastructure that enables business users to author, - in a controlled natural language -, execute and manage business rules in a collaborative way. As the majority of BRMS, it uses an object oriented models to formalize the domain knowledge. In WODM, this object oriented model is called BOM (Business Object Model). The BOM represents the entities of a given business (*i.e. patient, age*). It is generated over the XOM (eXecutable Object Model) then verbalized. The XOM is the model enabling the execution of rules. It references the application objects and data, and is the base implementation of the BOM. The XOM can be built from compiled Java classes (Java execution object model) or XML Schema (dynamic execution object model). The verbalization of the BOM consists of generating a controlled natural language vocabulary (VOC) which enables authoring the business rules (*i.e. the patient, the age of the patient*).

[8] http://www.w3.org/2004/OWL/

[9] http://www.aphp.fr

[10] This work is partially founded by the European Commission under the project ONTORULE (IST-2009-231875).

[11] http://www-01.ibm.com/software/decision-management/operational-decision-management/websphere-operational-decision-management/

2.1 Authoring Business Rules over OWL Ontologies

To enable business users to author business rules, in a natural controlled language, we developed the WODM OWL plug-in. This plug-in exploits infrastructure offered by WODM to import OWL ontologies within it. The main component for authoring rules in WODM is the BOM. For this, we performed a mapping of OWL concepts (TBox) into the BOM. Thus, when we import an OWL ontology within WODM, the BOM is automatically generated and the functionalities offered by the BRMS can be used. The general idea of the mapping is: ontology concepts are mapped into BOM classes and the properties are mapped into attributes of the classes. Nevertheless, due to the difference of the power of expressiveness between OWL ontology and the BOM, there are some OWL construct that could not map into the BOM [3].

2.2 Executing Business Rules over OWL ontologies

The process of executing business rules in WODM consists of several steps. Business rules, authored in a controlled natural language are translated into executable rules, which are written in a formal technical rule language IRL (ILOG Rule Language). During this translation, the references to the BOM's classes and properties are translated to references into the XOM. When the input provided to WODM is a Java object model, the XOM is built from this model. But in our case, the input provided to WODM is an OWL model.

To execute business rules authored over ontologies, we perform a second mapping of OWL/BOM entities to a XOM using Jena. Jena is a Java framework, including an ontology API for handling OWL ontologies, which allows generating Java objects from the entities of the ontology. These Java objects then constitute the XOM. The use of Jena provides an execution layer for the OWL ontologies. This execution layer provides inference mechanisms on this model and the mapping of OWL concepts, properties, and individuals to a Java object model.

3 Experimentation

The method described above enables to author and execute business rules over OWL ontology. This method can be used in different business domain. In our case, to experiment our work, we used a pharmaceutical validation use case to implement clinical decision rules for pharmaceutical validation of medication orders.

To illustrate our work, we present a business scenario that stages three personas representing business users involved in the design and the implementation of a set of clinical decision rules for pharmaceutical validation of medication orders (see Figure 1).

Marc is the business analyst. His mission is to formalize the business knowledge and to make sure that the business model (i.e. ontology) is correct, complete

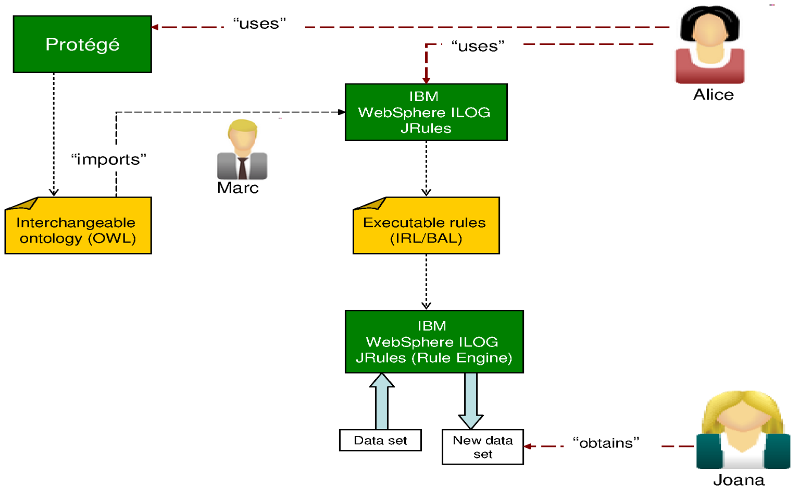

Protégé
"uses"
"uses"
Alice
IBM WebSphere ILOG JRules
"imports"
Marc
Interchangeable ontology (OWL)
Executable rules (IRL/BAL)
IBM WebSphere ILOG JRules (Rule Engine)
Data set
New data set
"obtains"
Joana

Fig. 1. Business scenario.

and valid. Alice is the domain expert. In this business scenario two domain experts interact which are a pharmacist and a physician. They understand the formalization of the rules and uses business rules tools. They are in charge of editing the clinical decision rules in the business application. Joana is the operational user who is the pharmacist. She uses the application to verify if a given prescription is valid or not.

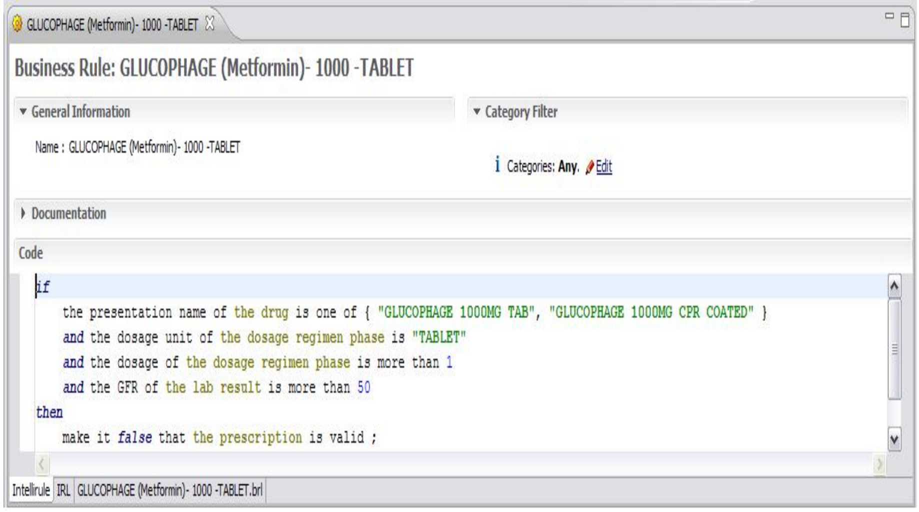
The ontology used to build this business application and the authored business rules that we will show in the following have been built based on the work presented in [2]. This work has been made in collaboration with pharmacists and physicians from the Georges Pompidou University Hospital [4]. The ontology is composed of 17 concepts and 25 properties. We will focus on 5 concepts and their properties used to author the rules we present in this paper. The ontology contains a concept **Patient** which has **LabResult** and is concerned by a **Prescription** that has a **DosageRegimenPhase**. The rules authored over this ontology test on the *presentation name* of a **Drug**, the *dosage unit* and the *dosage* of the **DosageRegimenPhase** of a **Prescription** and on the *GFR* (Glomerular Filtration Rate) of the **LabResult** of a **Patient**. Depending on the values given to these properties, they assign if a **Prescription** is *valid* or not.

When Marc finish the edition of the ontology using an ontology editor (i.e. Protégé), he imports it into WODM which automatically generate the BOM. Once the BOM is generated, the domain experts (pharmacists and physicians) author the clinical decision rules in natural controlled language. Two examples of authored rules are presented in the following (see Figure 2 & 3).

The rule in Figure 2 tests if :

- the presentation name of a drug is “GLUCOPHAGE 1000MG TAB” or “GLUCOPHAGE 1000MG CPR COATED”
- the dosage unit of the dosage regimen phase is “TABLET”
- the dosage of the dosage regimen phase is more than 1
- the GFR of the lab result of a patient is more than 50

Then it sets that the prescription is not valid.



Business Rule: GLUCOPHAGE (Metformin)- 1000 -TABLET

Name : GLUCOPHAGE (Metformin)- 1000 -TABLET

Categories: Any.

```
if
    the presentation name of the drug is one of { "GLUCOPHAGE 1000MG TAB", "GLUCOPHAGE 1000MG CPR COATED" }
    and the dosage unit of the dosage regimen phase is "TABLET"
    and the dosage of the dosage regimen phase is more than 1
    and the GFR of the lab result is more than 50
then
    make it false that the prescription is valid ;
```

Fig. 2. GLUCOPHAGE (Metformin)-1000.

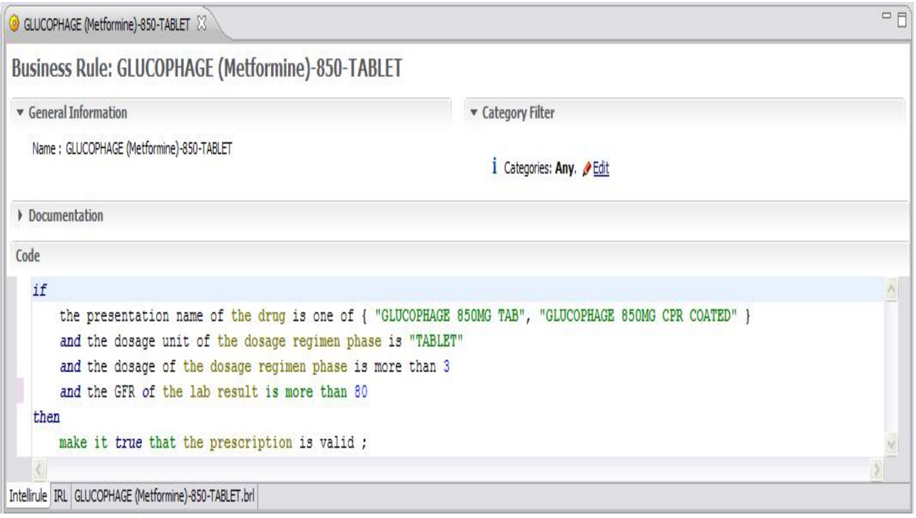
The rule in Figure 3 tests if:

- the presentation name of a drug is “GLUCOPHAGE 850MG TAB” or “GLUCOPHAGE 850MG CPR COATED”
- the dosage unit of the dosage regimen phase is “TABLET”
- the dosage of the dosage regimen phase is more than 3
- the GFR of the lab result of a patient is more than 80

Then it sets that the prescription is valid.

The pharmacist enters the data concerning a prescription and launches the execution of the rules which will determine if the prescription is valid or not. For example, Joana enters data concerning two prescriptions given to two different patients (see Table 1).

Prescription 1 for patient 1 who has the GFR of his lab result equals to 90. The dosage unit of the dosage regimen phase of the prescription is TABLET and its dosage is 2. The prescription contains a drug called GLUCOPHAGE 1000MG TAB. In this case the rule called GLUCOPHAGE - 1000 TABLET Rule (see Figure 2) will be launched and set the validation of the prescription to false.



GLUCOPHAGE (Metformine)-850-TABLET
Business Rule: GLUCOPHAGE (Metformine)-850-TABLET
General Information
Category Filter
Documentation
Code
Intellirule
IRL
GLUCOPHAGE (Metformine)-850-TABLET.brl

Name : GLUCOPHAGE (Metformine)-850-TABLET

Categories: Any. Edit

```
if
    the presentation name of the drug is one of { "GLUCOPHAGE 850MG TAB", "GLUCOPHAGE 850MG CPR COATED" }
    and the dosage unit of the dosage regimen phase is "TABLET"
    and the dosage of the dosage regimen phase is more than 3
    and the GFR of the lab result is more than 80
then
    make it true that the prescription is valid ;
```

Fig. 3. GLUCOPHAGE (Metformin)- 850.

Prescription 2 for patient 2 who has the GFR of his lab result equals to 95. The dosage unit of the dosage regimen phase of this prescription is TABLET and its dosage is 4. The prescription contains a drug called GLUCOPHAGE 850 MG TAB. In this case the rule called GLUCOPHAGE (Metformin) 850 (see Figure 3) will be launched and the validation of the prescription will be set to true.

Table 1. Prescriptions of the pharmacist

	Presentation name	Dosage unit	Dosage	GFR
P1	GLUCOPHAGE 1000MG TAB	TABLET	2	90
P2	GLUCOPHAGE 850 MG TAB	TABLET	4	95

4 Conclusion

In this paper, we present an application of pharmaceutical validation of medication orders that implement clinical decision rules, in a natural controlled language, over an OWL ontology. In order to develop this application we use the prototype described in Section 2 that enables authoring and executing business rules. The clinical decision rules designed in this study will be integrated with the HEGP clinical information system as an alert system for more assessment.

In perspective, we propose to improve the rule presented in this work in order to have recommendations that enable to revise the invalid prescription. Such recommendations, considered as alerts, inform the pharmacist about the cause of the invalidity of a prescription.

One particularity of ontologies is that they evolve over time. Ontology evolutions consist of changes that could impact an ontology. The business rules depend on the entities of the ontology and its evolution may have an impact on them and causes inconsistencies. This is an issue on which we focus and for which we developed the $\mathcal{MDR}$ approach (Model-Detect-Repair) [5]. This approach enables to tracks ontology changes, detects the rule inconsistencies that could be caused by a change and then proposes solution, called repair, to repair the inconsistencies.

References

1. G. Antoniou, C. V. Damasio, B. Grosof, I. Horrocks, M. Kifer, J. Maluszynski, and P. F. Patel-Schneider. Combining rules and ontologies: A survey. *Technical Report IST506779/Linkoping/I3-D3/D/PU/a1, Linkoping University*, 2004. http://rewerse.net/publications/.
2. A. Boussadi, C. Bousquet, B. Sabatier, T. Caruba, P. Durieux, and P. Degoulet. A business rules design framework for a pharmaceutical validation and alert system. *Methods of Information in Medicine*, 2011.
3. A. Chniti, S. Dehors, P. Albert, and J. Charlet. Authoring business rules grounded in owl ontologies. In M. Dean et al. (Eds.), editor, *RuleML 2010 : The 4th International Web Rule Symposium: Research Based and Industry Focused.* LNCS 6403, Springer-Verlag Berlin Heidelberg 2010, 2010.
4. P. Degoulet, L. Marin, M. Lavril, C. Le Bozec, E. Delbecke, J-J. Meaux, and L. Rose. The hegp component-based clinical information system. *. Int J Med Inform; 69(2-3):*, pages 115–126, 2003.
5. M. Fink, A. El Ghali, A. Chniti, R. Korf, A. Schwichtenberg, F. Lévy, J. Pührer, and T. Eiter. D2.6 consistency maintenance. final report. *ONTORULE Delivrable, http://ontorule-project.eu/deliverables.*, 2011.
6. V. Kashyap, A. Morales, and T. Hongsermeier. On implementing clinical decision support: Achieving scalability and maintainability by combining business rules and ontologies. *AMIA Annu Symp Proc*, pages 414–418, 2006.
7. Y. Kawazoe and K. Ohe. An ontology-based mediator of clinical information for decision support systems: a prototype of a clinical alert system for prescription. *Methods Inf Med, vol. 47, no. 6*, pages 549–559, 2008.

Keynote

How clinical and genomic data integration can support pharmacogenomics efforts related to personalized medicine

Eric Neumann, Clinical Semantics Group

Abstract*:* Pharmacogenomics is a key factor that will drive the personalized medicine vision. It will help create new combinations of clinical and genomic information necessary for making clinical decision in personalized medicine. Much can be done using available public data sources, yet they lack essential facts that can be only obtained from deep focused curation. A clear distinction will be made between Data vs Actionable Knowledge, whereby semantically linked data specifically can be leveraged to support decision making for personalized medicine.

Keynote

Understanding Recovery as a Mechanism for Individualized Treatment Selection in Major Depressive Disorder: A case study

Joanne S. Luciano, Tetherless World Constellation @ Rensselaer Polytechnic Institute

Abstract*:* Depression is a mental disorder, characterized by symptoms of sadness, loss of interest or pleasure, feelings of guilt or low self-worth, disturbed sleep or appetite, feelings of tiredness, and poor concentration. The World Health Organization (WHO) reports "Depression affects more than 350 million people of all ages, in all communities, and is a significant contributor to the global burden of disease." The US government reports that Major Depressive Disorder (MDD) is the leading cause of disability in the U.S. for ages 15-44. This talk will present a model of depression recovery used to characterize individual patient response to treatment. The model provides an explanation that on the surface seems like a paradox, namely, how an antidepressant treatment could result in suicide. The talk will be placed in the context of the past twenty years and highlight key events that are leading to radically different world of medical practice. I'll briefly mention some current controversies in the treatment of Depression and introduce the emerging field of Health Web Science.