3rd International Workshop on Linked Science 2013 - Supporting Reproducibility, Scientific Investigations and Experiments

Workshop in conjunction with the
12th International Semantic Web Conference 2013
Sydney, Australia, 21 October 2013

Edited by:
Paul Groth
Marieke van Erp
Tomi Kauppinen
Jun Zhao
Carsten Keßler
Line C. Pouchard
Carole Goble
Yolanda Gil
Jacco van Ossenbruggen

Preface

The 3rd edition of the Linked Science (LISC 2013) workshop (http://linkedscience.org/events/lisc2013/) was hosted at the International Semantic Web Conference. Like prior workshops in the series, this Linked Science had a theme. It focused on using semantic technologies to represent data and methods and enable their knowledge discovery, reuse and validation. The program was divided into two parts - traditional presentations of papers and an interactive discussion and co-writing session which resulted in a set of challenges to the research community. There were 26 attendees.

The paper session consisted of a keynote by Prof. Carole Goble of the University of Manchester titled "Results may vary: reproducibility, open science and all that jazz." In addition, there were 7 presentations covering topics ranging from reproducing a pharmacovigilance case study to capturing intent behind scientific experiments. A key theme was tackling specific scientific problems by combining existing techniques. We hope you enjoy reading this set of state-of-the art research.

As with prior workshops, Linked Science 2013 continued the tradition of "working" at the workshop. The attendees were divided into three groups and asked to to develop matrices about how semantic web/linked data solutions can help address reproducibility/re* problems. These matrices were then presented by group leaders, which were filmed. The group then developed a series of challenges to the Linked Data community with respect to addressing these re* problems. The challenges were:

1. Promote the basics of linked data for reproducibility
2. Integrate Semantic Web technologies and the publishing process.
3. Make it easier to publish data and then work with it than work directly on your own data.
4. Provide an integrated view of the how, what, when, where, and why of the scientific process.
5. Provide a mechanisms for dealing with copyright on data both from a technical and social perspective.
6. Get an altmetric based award into one of our own venues.
7. Make sure the EBI RDF platform does not get shut down in two years.

We hope these challenges will spur thinking in the community.

The challenges as well as the videos and matrices were all made available on the Figshare data sharing service and can be cited as:

Groth, Paul; Ansell, Peter; Kjernsmo, Kjetil; Van Ossenbruggen, Jacco; Palma, Guillermo; Goble, Carol; Vidal, Maria-Esther; McLean, Cameron; Hosking, Richard; Cassidy, Steve; Zhao, Jun; Gupta, Prashant; Ockeloen, Niels; Klyne, Graham (2013): LISC 2013 - Results: Discussion Groups on Semantic Web and Reproducibility. figshare. `http://dx.doi.org/10.6084/m9.figshare.828798`

Overall, this edition continued providing a successful forum for discussing how semantic web technologies and linked data can help science.

We wanted to thank the entire program committee for helping to assemble the program and the attendees for their enthusiastic participation.

- The LISC 2013 Co-organizers:
Paul Groth, VU University Amsterdam
Marieke van Erp, VU University Amsterdam
Tomi Kauppinen, Aalto University
Jun Zhao, University of Oxford
Carsten Keßler, University of Muenster
Line C. Pouchard, US Department of Energy
Carole Goble, University of Manchester
Yolanda Gil, Information Sciences Institute
Jacco van Ossenbruggen, VU University Amsterdam

Programme Committee

The following colleagues kindly served in the workshop's program committee. Their joint expertise covers all of the questions addressed in the workshop, and they reflect the range of relevant scientific communities.

- Mathieu d'Aquin, The Open University, UK
- Charalampos Bratsas, Aristotle University of Thessaloniki, Greece
- Boyan Brodaric, Natural Resources Canada, Canada
- Arne Bröring, 52° North, Germany
- Gully Burns, ISI, University of Southern California
- Oscar Corcho, UPM, Spain
- Stefan Dietze, L3S Research Center, Germany
- Ying Ding, Indiana University, USA
- Hannes Ebner, Royal Institute of Technology (KTH), Sweden

- Antske Fokkens, VU University, Amsterdam
- Asunción Gómez Pérez, Universidad Politécnica de Madrid
- Willem van Hage, Vrije Universiteit, the Netherlands
- Frank van Harmelen, VU University, Amsterdam
- Michiel Hildebrand, VU University, Amsterdam
- Rinke Hoekstra, Vrije Universiteit, the Netherlands
- Laura Hollink, VU University, Amsterdam
- Michael D. Huhns, University of South Carolina, USA
- Stratos Idreos, CWI, Amsterdam
- Krzysztof Janowicz, University of California, Santa Barbara, USA
- Craig A. Knoblock, University of Southern California, USA
- Werner Kuhn, University of Muenster, Germany
- Timothy Lebo, RPI, Troy, NY, USA
- Zoltn Mikls, University of Rennes 1, France
- Paulo Pinheiro da Silva, Pacific Northwest National Laboratory, USA
- Herbert van de Sompel, Los Alamos National Laboratory, USA
- Eric Stephan, Pacific Northwest National Laboratory, USA
- Mark Wilkinson, Centre for Plant Biotechnology and Genomics UPM-INIA, Madrid, Spain
- Bryn Williams-Jones, Connected Discovery & OpenPHACTS, UK
- Max Wilson, University of Nottingham, UK
- Amrapali Zaveri, University of Leipzig, Germany

Contents

Results May Vary: Reproducibility, Open Science and All That Jazz

Carole Goble

School of Computer Science
University of Manchester

Abstract. How could we evaluate research and researchers? Reproducibility underpins the scientific method: at least in principle if not practice. The willing exchange of results and the transparent conduct of research can only be expected up to a point in a competitive environment. Contributions to science are acknowledged, but not if the credit is for data curation or software. From a bioinformatics view point, how far could our results be reproducible before the pain is just too high? Is open science a dangerous, utopian vision or a legitimate, feasible expectation? How do we move bioinformatics from one where results are post-hoc made reproducible, to pre-hoc born reproducible? And why, in our computational information age, do we communicate results through fragmented, fixed documents rather than cohesive, versioned releases? In this talk, which I gave as a keynote at the 2013 joint conference Intelligent Systems in Molecular Biology / European Conference on Computational Biology, I will explore these questions drawing on 20 years of experience in both the development of technical infrastructure for Life Science and the social infrastructure in which Life Science operates.

Building Executable Biological Pathway Models Automatically from BioPAX

Timo Willemsen, Anton Feenstra, and Paul Groth

Department of Computer Science, VU University Amsterdam, The Netherlands

timo.willemsen@gmail.com,{k.a.feenstra,p.t.groth}@vu.nl

Abstract. The amount of biological data exposed in semantic formats is steadily increasing. In particular, pathway information (a model of how molecules interact within a cell) from databases such as KEGG and WikiPathways are available in a standard RDF-based format BioPAX. However, these models are *descriptive* and not *executable* in nature. Being able to simulate or execute a pathway is one key mechanism for understanding the operation of a cell. The creation of executable models can take a significant amount of time and only relatively few such models currently exist. In this paper, we leverage the availability of semantically represented pathways, to bootstrap the creation of executable pathway models. We present an approach to automate the creation of executable models in the form of Petri-Nets from BioPAX represented pathways. This approach is encapsulated in an online tool, BioPax2PNML.

Keywords: biological pathways, biological networks, BioPax, executable models, Petri nets

1 Introduction

A biological pathway, simply said, is a sequence of interactions among molecules of a cell. There are many different types of pathways; gene regulation pathways, signaling pathways and protein interaction pathways are among the most commonly used ones. [1]

Originally, pathways were hand-drawn and presented in papers. Pathways are now made available in online databases in computer parsable formats (e.g. BioPAX). For example, the WikiPathways has over 1700 available pathways[1]. While these pathway descriptions are highly useful, they contain mostly static information about interacting molecules and do not describe how *pathways actually work* or give insight into the dynamics of these interactions [2].

To address this lack of information, work has been undertaken to create computational models of these pathways [3]. Two types of models can be distinguished: executable and mathematical [4]. The mathematical models give insight into quantities and how they change over time, and are frequently created by systems biologists. Executable models are valuable to biologists because they have

[1] See http://WikiPathways.org/index.php/WikiPathways:Statistics for statistics on WikiPathways

a large variety of uses [4,5]. They can be used to summarize available knowledge of interactions and mechanisms in a system, and to investigate how components cooperate to produce global system behaviour. Creating an executable model is still a tedious manual process, mostly because they contain parameters that need to be collected manually. On the other hand, mathematical models typically require detailed knowledge of (kinetic and rate) parameters, which are often not available and can be very hard to obtain from experiments. From our experience, for executable models, the process of model construction and parameter calibration usually takes several months [3,6,7], even for a modestly sized network. This is currently one of the major bottlenecks in computational life sciences research [8].

This paper begins to address this bottleneck by leveraging the availability of semantic representations of pathways and converting them to an executable model. Concretely, the contributions of this paper are: *i)* to present a method to automate validation of pathway data; *ii)* a mapping of the BioPAX format to an executable model (Petri nets, represented in the Petri Net Markup Language; PNML); and, *iii)* a method to automatically create these executable models. We have developed a webservice that encapsulates the described method and can be accessed at www.few.vu.nl/∼twn370/BioPax2PNML/. Additionally, all code is available online at: https://github.com/TimoWillemsen/Biopax2PNML.

The rest of this paper is organized as follows. We begin in Section 2 with background information on biological pathways and common formats for both descriptive (BioPAX) and executable (PNML) representations of them. We then describe our approach for mapping between these two formats (Section 3). To ensure that a BioPAX pathway has the appropriate information to be converted to PNML, we present a validation approach in Section 4. This is followed, in Section 5, by a description of the implementation of our method. Finally, we conclude with some thoughts on future work in Section 6.

2 Biological Pathways

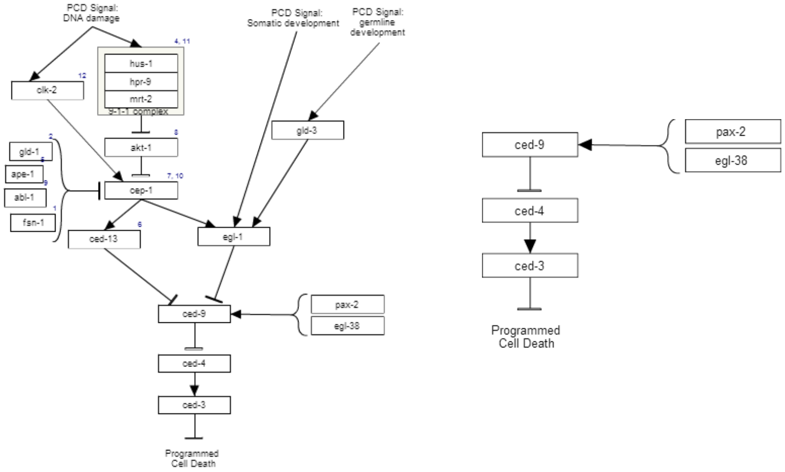
There are different types of biological pathways, corresponding to different levels of abstraction. For example, a pathway may describe interactions between different cells, or between genes, or between proteins, or it may describe biochemical reactions (or combinations thereof). Many databases exist that collect this information in a variety of forms, and some are very specialized on particular types of data. It is beyond the scope of this work to provide a comprehensive overiew. Some of the most well-known are WikiPathways [9], focused on signal transduction; the KEGG Pathway database [10,11], with a focus on metabolic pathways; and Reactome [12] which has a broader scope.

The examples provided in this paper will focus on signal transduction pathways, as these tend to be well-studied and therefore well-defined. Such pathways typically include protein-protein interactions, protein-gene interactions and biochemical reactions.

We have based our research on the pathways provided by the WikiPathways database [9]. This is a community-driven service where biological pathways are extensively manually curated. The context of the pathways included in WikiPathways can vary considerably, depending on their intended use. For example, simply representing known interactions in a shareable way is considered useful, but such pathways likely will not include details that are crucial for computational analysis, even as simple as explicit notation of interactions among proteins and genes. As a result of this, only certain pathways are suitable for computational analysis.

One such example is the *C. elegans* Programmed Cell Death pathway from the WikiPathways Database, as shown in Fig. 1.

Fig. 1. *C. elegans* Programmed Cell Death Pathway from the WikiPathways Database ID:WP367. The left panel shows the complete pathway, the right panel shows the subset of 5-genes used.



For the purpose of this paper, we have taken a subset of this pathway, as shown in Fig. 1. This pathway consists of 5 genes. When ced-3 is activated, it will trigger the cell's programmed death.

We now discuss the computational representation of pathways used by WikiPathways. After which, we briefly describe the use of Petri-nets to as a language for executable models of pathways.

2.1 BioPax

In 2010 Demir et al [13] created the Web Ontology Language (OWL) based standard for modeling pathways: BioPax. A key aspect of this standard is that it allows for referring to external databases for information (e.g. linking to UniProt protein descriptions.) This standard has been used in many different biological databases; all the three mentioned above, Reactome, KEGG and WikiPathways expose BioPax through an RDF interface [9,11,12].

BioPax can be used to model different types of pathway components. An example of how genes are modelled in BioPax, is shown below; the ced-3 and ced-4 genes of the *C. elegans* Programmed Cell Death pathway, as shown in Fig. 1.

Two genes, ced-3 and ced-4, from the C. elegans Programmed Cell Death Pathway from the WikiPathways Database `ID:WP367`

```
<bp:Protein rdf:about="eef1e">
 <bp:displayName>ced-3</bp:displayName>
 <bp:entityReference rdf:resource="id3" />
</bp:Protein>
<bp:Protein rdf:about="c0b3e">
 <bp:displayName>ced-4</bp:displayName>
 <bp:entityReference rdf:resource="id4" />
</bp:Protein>
```

An example of interactions in a pathway modelled in BioPax is shown below; we see a reaction '`id40`' that connects a right-hand-side element (`eef1e`; ced-3) with a left-hand-side (`c0b3e`; ced-4) element.

Gene interaction of the C. elegans Programmed Cell Death Pathway from the WikiPathways Database `ID:WP367`

```
<bp:BiochemicalReaction rdf:about="id40">
 <bp:right rdf:resource="eef1e" />
 <bp:left rdf:resource="c0b3e" />
</bp:BiochemicalReaction>
```

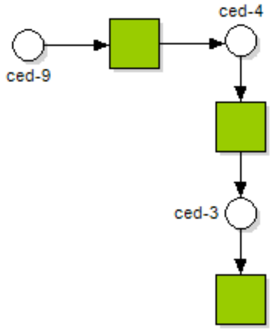
2.2 Petri nets

Petri nets are a formalism geared towards modelling and analysis of concurrent systems. A Place-Transition (PT) Petri net is a quadruple (P, T, A, m), where P is a set of places and T a set of transitions. A describes arcs which connect places with transitions or vice versa. Each place holds zero or more tokens, which represent flow of control through this place. The number of tokens in each place all together are called a marking m of the network.

Fig. 2 shows a graphical representation of such a Petri Net, again for our small example part of the *C. elegans* Programmed Cell Death pathway. Squares are transitions, representing interactions, and circles are places, representing genes. Arcs are represented by arrows, and the marking is empty. Firing of a transition depends on the availability of resources (tokens) in the input places, and represents the execution of a reaction: consuming substrates and creating products.[14,15]

For computational purposes we have chosen to represent Petri nets in the Petri Net Markup Language (PNML) format. This is a straightforward XML

Fig. 2. An example Petri net of a small part of the *C. elegans* Programmed Cell Death Pathway (WikiPathways:WP367)



standard that a number of systems support.[16] Fig. 2.2 shows the Petri net of Fig. 2 in an XML representation. Petri nets are recognized as a powerful tool to model biological pathways [14,15], as the formalism readily allows to capture both the complexity and the highly concurrent nature of biological systems, while optimally leveraging the large amounts of qualitative data available.[15,3,6]

Fig. 3. *PNML representation of the C. elegans Programmed Cell Death pathway (WikiPathways:WP367) Petri net as shown in Fig. 2.*

```
<transition id="t11">
</transition>
<place id="eef1e">
    <name>
        <text>ced-3</text>
    </name>
</place>
<place id="c0b3e">
    <name>
        <text>ced-4</text>
    </name>
</place>
<arc id="a2" source="c0b3e" target="t11" />
<arc id="a3" source="t11" target="eef1e" />
```

3 BioPax to PNML mapping

To transform static BioPax data into an executable Petri net, we have developed a mapping between the two formats. BioPax is an RDF format, while PNML is an XML format. It should be taken into account that the semantic linking is lost when a BioPax pathway is converted to PNML Petri-net. For example, genes or proteins have different identifiers in different databases. BioPax gives a way to link multiple identifiers to a gene or protein, but PNML does not support this feature.

3.1 Genes or Proteins

Each gene or protein is modelled as a place in the Petri net. Because the creation of the Pathways in WikiPathways has been done manually, often they are not consistent and may, for example, contain multiple instances of one gene or protein. The mapping does not take into consideration the fact that duplicate genes or proteins may represent the same entity and are modelled twice simply for readability, or rather that they are modelled twice because they represent a different entity of the same gene/protein (for example in a different location, or in a different state). However we address this issue with the validation rules introduced in Section 4.

The first stage in mapping is shown in Algorithm 1, which transforms BioPax proteins/genes to PNML.

Algorithm 1 Genes/Proteins BioPax to PNML

```
$P = \emptyset$
for all <bp:Protein> p in BioPax do
    if $p \notin P$ and p is other entity then
        add p to P
    end if
end for
```

3.2 Interactions

Interactions are also mapped to PNML. Each <bp:BiochemicalInteraction> is mapped to a transition. Then for each <bp:Left> an arc is added pointing into the transition and out from the corresponding place; for each <bp:Right> an arc is created pointing out of the transition and into the corresponding place. Algorithm 2 shows the straightforward way to do this.

Once both algorithms 1 and 2 are executed a Petri net is created. Formally, the Petri net can be described as $PN = P, T, A, \emptyset$ where P are the places, T the transitions, A the arcs and markings $m = \emptyset$ since there are no tokens in the system yet. In terms of modelling the biological system, the places represent biological entities, like genes, proteins or complexes, the transitions represent biochemical reactions and interactions, and the arcs represent the associations between these two. Tokens represent the availability of the resources of the corresponding place in the Petri net.

Algorithm 2 Gene/protein interaction BioPax to PNML

```
$T = \emptyset$
$A = \emptyset$
for all <bp:BiochemicalInteraction> t in BioPax do
    Add t to T
    for all <bp:Left> left in BioPax do
        left.in = t
        left.out = left.resource
        Add left to A
    end for
    for all <bp:Right> right in BioPax do
        right.in = right.resource
        right.out = t
        Add right to A
    end for
end for
```

If we then execute both Algorithm 1 and Algorithm 2 on Fig. 1, a petri net is generated. Part of the output is shown in Fig. 2.2

4 BioPax Validation

The mapping described in Section 3 is based on several assumptions about the contents of the input BioPax file. The basic assumptions are that genes, proteins and complexes (bound combinations of proteins, possibly including a gene) are entities, and that these entities can change state or identity only through biochemical interactions.

However, because of the manual nature of pathway construction, these assumptions may not hold for a given pathway instance in the database. To make sure the data is presented as it should be, we have developed a set of validation rules and a validator available online.

We have developed two types of validation rules; semantic and syntactic. The syntactic validation consists of basic RDF-validation. This is necessarily because from our preliminary survey, a large fraction of pathways are not modelled correctly for translation.

More interesting is the semantic validation. These rules ensure that the information contained in the model is consistent and complete enough to create an executable Petri net. Table 1 shows these validation rules.

These rules ensure that the provided BioPax file contains everything needed. We have categorized the validation rules by severity:

- **Category error** rules are minimal requirements for mapping.
- **Category warning** rules that mean the mapping can proceed but may lead to an unconnected or incomplete Petri net.

This framework is set up in a modular fashion, so that extension is easy.

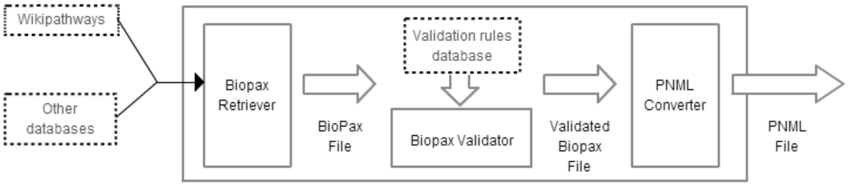
Table 1. Semantic Validation Rules

Id	Severity	Rule
1	Error	Each BioChemicalReaction should have a Left child element.
2	Error	Each BioChemicalReaction should have a Right child element.
3	Error	Each Pathway should have one or more PathwayComponents of type BiochemicalReactions.
4	Warning	Each BiochemicalReaction Left child is the actor of the interaction.
5	Warning	Each BiochemicalReaction Right child is the actant of the interaction.
6	Warning	Each unique entity of a protein/gene is modelled as a different Protein.
7	Warning	Each Protein should have a corresponding RelationshipXref.
8	Warning	Whenever a BiochemicalReaction has multiple Left or Right tags, it means that it has effect on multiple genes/proteins.
8	Warning	Protein complexes are modelled as a Complex tag.

5 Implementation

We have implemented the methods described above as a webservice. The service consists of 4 components: a validation rule database, a validator, a BioPax to PNML converter and a pathway retriever, as is shown schematically in Fig. 4.

Fig. 4. BioPax to PNML webservice architecture



5.1 Pathway retriever

The webservice provides an interface to query different datasources. At the time of writing only an interface to WikiPathways is provided, using the available webservices [17]. However, support for other generic BioPax could be a future extension.

The retriever queries WikiPathways and downloads the pathway in the BioPax format, so validation and conversion can be done.

5.2 Validation rule database

The validation rule database is a set of SPARQL queries. Each query returns a set of RDF triples that violate the rule (this set may be empty). This way feedback can be given about where the rule violation takes place in the BioPax File.

The way the database is set up allows easy addition of rules. This modularity makes it possible to improve on the current validation rules, but also allows validation rule sets for different types of pathways (for example signalling pathways vs. gene regulatory networks). Fig. 5 shows as an example the implementation of `rule 1` of Table 1.

Fig. 5. SPARQL *implementation of rule 1 of Table 1*

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX bp: <http://www.biopax.org/release/biopax-level3.owl#>

SELECT ?reaction
WHERE {
  ?reaction rdf:type bp:BiochemicalReaction.
    OPTIONAL {
        ?reaction bp:left ?left.
    }
    FILTER (!BOUND(?left))
}
```

This query returns every `bp:BiochemicalReaction` that does not have a `bp:left` child element associated to it.

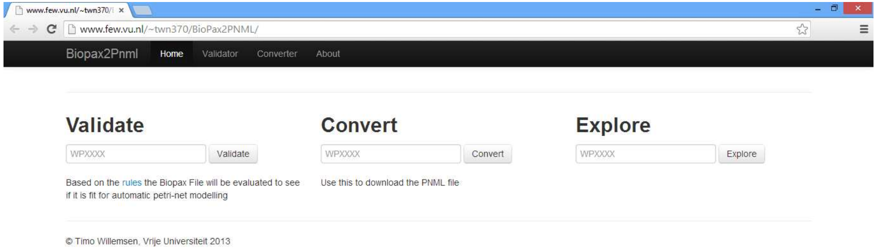
5.3 BioPax Validator

The biopax validator is software that can analyze BioPax files according to the validation rules provided by the rule database. It is essentially a graphical user interface around the SPARQL queries. It annotates the place where errors or warnings have occurred and provides an easy to use interface to solve them.

5.4 BioPax to PNML Converter

Once a BioPax file has been validated, the BioPax to PNML converter can be used to generate an executable Petri net. This converter works according to the mapping described in Section 3. This is implemented as an online tool, named BioPax2PNML, and can be accessed on www.few.vu.nl/∼twn370/BioPax2PNML/.

Fig. 6. *Screen shot of the user interface of the BioPax2PNML tool at* www.few.vu.nl/∼twn370/BioPax2PNML/.



Validate

Based on the rules the Biopax File will be evaluated to see if it is fit for automatic petri-net modelling

Convert

Use this to download the PNML file

Explore

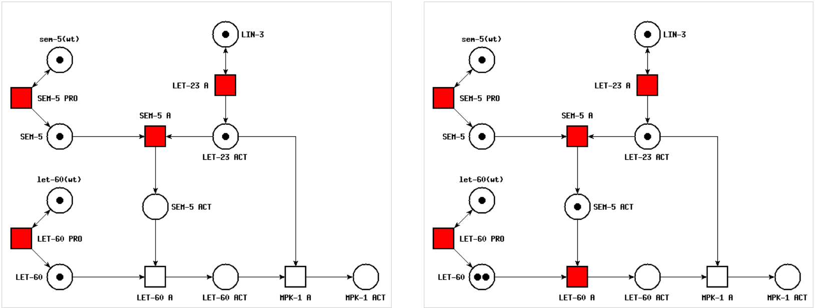
© Timo Willemsen, Vrije Universiteit 2013

5.5 Executing the PNML file

Although the proof of concept of the current work stops with the generation of a valid Petri net model in the form of a PNML file, it is nevertheless instructive to consider what subsequent steps should be. Execution of a Petri net can be performed under different execution semantics, however the most relevant for biological systems is commonly thought to be the so-called 'bounded asynchronous' execution [18,3,15]. Under this semantics, as many transitions as possible are executed simultaneously in each execution step. This represents the inherent concurrency of biological systems, where molecules typically act independently, certainly if they reside in different locations. This is also known as the 'token game', because execution of transitions has the effect of shifting tokens around the Petri net. Fig. 7 shows an example network and the change in state due to execution of enabled transitions.

Execution leads to a trajectory of markings, that represent the progression of states of the system in response to the intial marking, which corresponds to

Fig. 7. *Example of execution of a slightly non-trivial example network, taken from [15]. Enabled transitions (with input requirements satisfied; marked in red) will execute each step, execution of enabled transitions in the left panel will lead to the state shown in the right panel.*



a particular state or condition of the biological system. Typically, token levels are collected from a few places of interest and compared to experimental data of the corresponding biological molecule, or used to predict the behaviour of that particular molecule under the conditions modeled. Examples of these for signalling pathways can be found in [15,6], and for gene regulatory networks in [7].

6 Conclusion

Automatic Petri net creation of biological pathways is still a tedious process. The manual labor involved makes it so that even a modestly sized model can take several months to develop. In this paper we have provided a method to bootstrap this process. By using a mapping between the commonly used BioPax format and the PNML format, we have developed a way to automate the construction of Petri net models. Because biological information online may be inconsistent or incomplete, we have developed a set of validation rules to make sure that the data is suitable for automatic conversion.

To facilitate this, and as a proof of concept, an online tool BioPax2PNML that executes this and provide an easy interface for Petri net modelers to bootstrap the process of model creation.

The approach outlined here is an initial start to making fully developed executable models. In particular, deriving the weights on edges of the Petri nets is a challenging task. In terms of future work, we believe that by leveraging the links to other databases (e.g. Uniprot) we may be able to find additional information to infer such edge weights. Moreover, we may be able to connect additional parts of the resulting Petri-nets based on background knowledge about interactions contained in other databases or even use knowledge of chemistry provided by other data sources to create more precise models. A key foundation for work going forward is that Linked Data and Semantic Web standards facilitate the merging and acquisition of this information.

7 References

1. Ganesh A Viswanathan, Jeremy Seto, Sonali Patil, German Nudelman, Stuart C Sealfon. Getting Started in Biological Pathway Construction and Analysis *PLoS Comput Biol* **4**(2): e16, 2008
2. Pinney JW, Westhead DR, McConkey GA. Using Petri Net tools to study properties and dynamics of biological systems. *J Am Med Inform Assoc.* **12**(2):181-99, 2005.
3. Nicola Bonzanni, Elzbieta Krepska, K. Anton Feenstra, Wan Fokkink, Thilo Kielmann, Henri Bal, and Jaap Heringa. Executing multicellular differentiation: Quantitative predictive modelling of *C. elegans* vulval development. *Bioinformatics* **25**, 2049–2056, 2009.
4. Jasmin Fisher and Tom Henzinger. Executable cell biology. *Nature Biotechnology* **25**(11):1239–1249, November 2007.
5. Aviv Regev and Ehud Shapiro. Cellular abstractions: Cells as computation. *Nature* **419**:343, September 2002.
6. Bonzanni, N., Zhang, N., Oliver, S.G. and Fisher, J. The role of proteasome-mediated proteolysis in modulating activity of potentially harmful transcription factor activity in Saccharomyces cerevisiae. *Bioinformatics* **27**: i282–i287, 2011.
7. Nicola Bonzanni, Abhishek Garg, K. Anton Feenstra, Sarah Kinston, Diego Miranda-Saavedra, Judith Schutte, Jaap Heringa, Ioannis Xenarios, Berthold Göttgens. Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model *Bioinformatics* in press (2013).
8. Susanna-Assunta Sansone, Philippe Rocca-Serra, Dawn Field, Eamonn Maguire, *et al.* Toward interoperable bioscience data *Nat Genet* **44**(2): 121–126, January 2012.
9. Thomas Kelder, Martijn P. van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R. Conklin, Chris T. Evelo, Alexander R. Pico. WikiPathways: building research communities on biological pathways *Nucleic Acids Res* **40**(Database issue): D1301–D1307, January 2012.
10. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **7**:27–30, 2000.
11. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **7**(Database issue):D354–D357, 2006.
12. Robin A. Haw, David Croft, Christina K. Yung, Nelson Ndegwa, Peter D'Eustachio, Henning Hermjakob, Lincoln D. Stein. The Reactome BioMart In *Database*, Oxford, October 2011.
13. Emek Demir, Michael P. Cary, Suzanne Paley, Ken Fukuda, *et al.* BioPAX – A community standard for pathway data sharing *Nat Biotechnol* **28**: 935–942, September 2010.
14. Elzbieta Krepska, Nicola Bonzanni, K. Anton Feenstra, Wan Fokkink, Thilo Kielmann, Henri Bal, and Jaap Heringa. Design issues for qualitative modelling of biological cells with Petri nets. In *Proc. FMSB'08*, **5054** *LNCS*, 48–62. Springer, June 2008.

15. Nicola Bonzanni, K. Anton Feenstra, Wan Fokkink and Elzbieta Krepska. What can Formal Methods bring to Systems Biology? In: *Proc. FM'09*, **5850** *LNCS*, 16–22. Springer, 2009.
16. Masao Nagasaki, Ayumu Saito, Atsushi Doi, Hiroshi Matsuno, and Satoru Miyano. *Using Cell Illustrator and Pathway Databases.* Springer, April 2009.
17. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, *et al.* Mining Biological Pathways Using WikiPathways Web Services. *PLoS ONE* **4**(7): e6447, 2009
18. Jasmin Fisher, Tom Henzinger, Maria Mateescu, and Nir Piterman. Bounded asynchrony: A biologically-inspired notion of concurrency. In *Proc. FMSB'08*, Cambridge, **5054** *LNCS* 17–32. Springer, June 2008.

Exploiting Semantics from Ontologies and Shared Annotations to Find Patterns in Annotated Linked Open Data

Guillermo Palma[1], Maria-Esther Vidal[1], Louiqa Raschid[2], and Andreas Thor[3]

[1] Universidad Simón Bolívar, Venezuela
[2] University of Maryland, USA
[3] University of Leipzig, Germany
gpalma@ldc.usb.ve, mvidal@ldc.usb.ve, louiqa@umiacs.umd.edu, thor@informatik.uni-leipzig.de

Abstract. Linked Open Data initiatives have made available a diversity of collections that domain experts have annotated with controlled vocabulary terms from ontologies. The challenge is to explore these rich and complex annotated datasets, together with the domain semantics captured within ontologies, to discover patterns of annotations across multiple concepts that may lead to potential discoveries. We identify an annotation signature between a pair of concepts based on shared annotations and ontological relatedness. Formally, an annotation signature is a partitioning of the edges that represent the relationships between shared annotations. A clustering algorithm named *AnnSigClustering* is proposed to generate annotation signatures. Evaluation results over drug and gene datasets demonstrate the effectiveness of using annotation signatures to find patterns.

1 Introduction

Ontologies are developed by domain experts to capture knowledge specific to some domain. They have been extensively developed and widely adopted in the last decade. Simultaneously, Linked Open Data initiatives have made available a diversity of collections that have been annotated with controlled vocabulary (CV) terms from these ontologies. For example, the biomedical community has taken the lead in such activities; every model organism database has genes and proteins that are widely annotated with CV terms from the Gene Ontology (GO). The NCI Thesaurus (NCIt) version 12.05d has 93,788 terms and the LinkedCT dataset of clinical trial results *circa* September 2011 includes 142,207 drugs or interventions, 167,012 conditions or diseases, and 166,890 links to DBPedia, DrugBank and Diseasome. At the opposite end of the domain spectrum, the Financial Industry Business Ontology (FIBO) captures knowledge about the structure, properties and behavior of financial contracts.

The challenge is to explore these rich and complex annotated datasets, together with the domain semantics captured within ontologies, to discover patterns of annotations across multiple concepts that may lead to potential discoveries. For genes, these patterns may involve cross-genome functional annotations, e.g., combining the GO functional annotations of two model organisms such as Arabidopsis thaliana (a plant) and C. elegans (a nematode or worm), to predict new gene function or protein-protein interactions. Drug target prediction, with a goal of finding new targets for existing drugs, has received widespread media attention and has resulted in some notable successes, e.g., `Viagra`. Additional applications include predicting potentially adverse side-effects or providing a comprehensive summary of drug effectiveness so that health professionals may find cost-effective treatments [10].

As a first step to discovering complex annotation patterns, we define an *annotation signature* between a pair of scientific concepts, e.g., a pair of drugs or a pair of genes. The annotation signature builds upon the shared annotations or shared CV terms between the pair of concepts. The signature further makes use of knowledge in the ontology to determine the ontological relatedness of the shared CV terms. The annotation signature is represented by N groups (clusters) of ontologically related shared CV terms. For example, the annotation signature for a (drug, drug) pair will be a set of N clusters, where each cluster includes a group of ontologically related disease terms from the NCIt.

Given a pair of concepts, and their sets of annotations, A_i and A_j from ontology O, elements $a_i \in A_i$ and $a_j \in A_j$ form the nodes of a bipartite graph BG. Between nodes a_i and a_j there may be an edge or a path through O; an edge is the special case where a_i and a_j are identical CV terms from O. There may be a choice of paths between a_i and a_j depending on the the ontology structure and relationship types captured within O. One can use a variety of similarity metrics, applied to the edges and paths through the ontology O, to induce a weighted edge between a_i and a_j in BG; the weight represents the (ontologically related) similarity score in the range $[0.0, 1.0]$ between a_i and a_j.

Our objective is to determine an annotation signature based on the bipartite graph BG. There are many alternatives to create the signature. One could partition the edges of BG with possible overlap of the nodes. Another solution is to cluster the nodes and edges of BG. One may also consider a one-to-one bipartite match [8].

We define a version of the *Annotation Signature Partition* problem as the partitioning of the edges of BG into clusters such that the value of the aggregated cluster density is maximized; we will define the density metric in the paper. We develop *AnnSigClustering*, a clustering solution that implements a greedy iterative algorithm to cluster the edges in BG. We note that such a clustering will result in N clusters of the edges of BG with potential overlap of nodes in different clusters.

We perform an extensive evaluation of the effectiveness of the annotation signature on real-world datasets of genes and their GO annotations, as well as on the LinkedCT dataset of drugs and diseases from NCIt and their associations through the clinical trials.

Our research focuses on exploiting domain specific semantic knowledge. This includes both the ontology structure and relationship types between concepts. We show that by using the ontology structure to tune the (ontologically related) similarity score between node pairs a_i and a_j, we can control the annotation signature to produce clusters of more closely related terms that are more useful to the domain scientist. Further, the choice of specific relationship types can be used to further refine the clusters of CV terms in the annotation signature.

The contributions of this paper can be summarized as follows: *i*) Definition of an *annotation signature* to mine annotated datasets together with domain specific semantic knowledge captured within ontologies. *ii*) A greedy iterative algorithm that exploits knowledge encoded in an ontology to discover the signature of a pair of annotated concepts. *iii*) An empirical study that suggests that annotation signatures represent interesting patterns across drugs and across genes.

This paper is organized as follows: Section 2 presents annotation graphs from different domains and Section 3 defines our approach. Experimental results are reported in Section 4, while related work is summarized in Section 5. Section 6 concludes.

2 Motivating Example

An antineoplastic agent is a substance that inhibits the maturation, growth or spread of tumor cells. Monoclonal antibodies that are also antineoplastic agents have become an important tool in cancer treatments. When used as a medication, the non-proprietary drug name ends in -mab. Scientists are interested in studying the relationships between drugs and the corresponding diseases; drugs are annotated with the NCIt terms that correspond to the conditions that have been tested for these drugs. Figure 1 illustrates `Brentuximab vedotin` and `Catumaxomab` and some of their annotations. Each path between a pair of conditions, e.g., `Colorectal Carcinoma` and `Stage IV Rectal Cancer` through the NCIt is identified using red ovals which represent CV terms from the NCIt. From Figure 1, we may conclude that the shared disease signature for this pair of drugs includes five components. The three terms `Colon Carcinoma`, `Colorectal Carcinoma` and `Stage IV Rectal Cancer` may form one component. Similarly, another component may include `Thyroid Gland Neoplasm`, `Oropharyngeal Neoplasm` and `Head and Neck Neoplasm`.

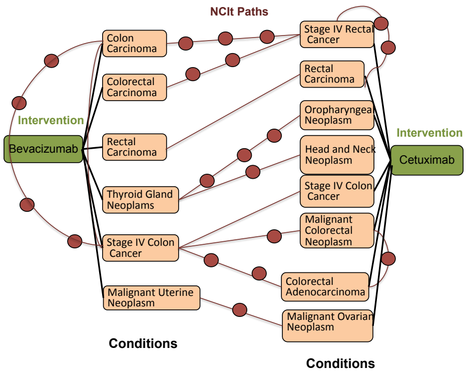


Fig. 1. Annotation graph representing the annotations of Brentuximab vedotin and Catumaxomab. Drugs are green rectangles; diseases are pink rectangles; NCIt terms are red ovals.

Consider a pair of financial contracts representing bonds (corporate, municipal, state, sovereign, etc.) from a repository such as EMMA [1]. Figure 2 shows an example of two bond contracts (green rectangles). These bonds are described by their CUSIP identifier, maturity date, principal, initial offering price, yield, etc. Each contract is also

[1] `http://www.emma.msrb.org/`

associated with a set of FIBO terms (pink ovals). For example, the Financial Contract A is associated with five terms including `Joint Guaranty` and `State Gurantor` while the Financial Contract is associated with seven terms. There is an edge with similarity equal to 1.0 between identical FIBO terms as well as paths through the FIBO ontology and intermediate FIBO terms (red circles).

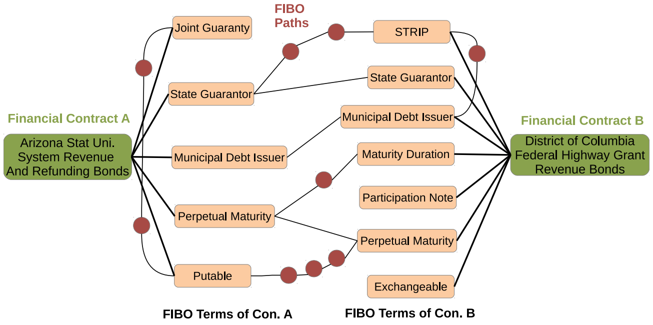


Fig. 2. FIBO terms (pink ovals) annotate a pair of financial contracts (green rectangles). An edge connects identical FIBO terms in the bipartite graph between the two sets of annotations on the left and right. Paths pass through intermediate FIBO terms (red circles).

3 Our Approach

A broad variety of similarity metrics have been proposed in the literature and have been summarized in [2]. Existing similarity metrics include the following: *i*) string-similarity metrics that measure similarity using (approximate) string matching functions; *ii*) path-similarity metrics such as *PathSim* and *HeteSim* that compute relatedness based on the paths that connect concepts in a graph; and *iii*) topological-similarity metrics that measure relatedness in terms of the closeness of CV terms in a given taxonomy or ontology.

We use a taxonomic distance metric d_{tax} [2]. The intuition behind the d_{tax} metric is to capture the taxonomic distance between two vertices with respect to the depth of the common ancestor of these two vertices. Additionally, d_{tax} tries to assign low(er) values of taxonomic distance to pairs of vertices that are: (1) at greater depth in the taxonomy and (2) are closer to their lowest common ancestor. A value close to 0.0 means that the two vertices are close to the leaves and both are close to their lowest common ancestor. A value close to 1.0 represents that both vertices are general or that the lowest common ancestor is close to the root of the taxonomy. Then, (1 - d_{tax}) will be used as the similarity or *ontological relatedness* between the two nodes.

The taxonomic distance metric d_{tax} is as follows, where *root* is the root node in the ontology; lca is the lowest common ancestor, and *pl* denotes path length:

$$d_{tax}(x, y) = \frac{pl(lca(x, y), x) + pl(lca(x, y), y)}{pl(root, x) + pl(root, y)} \quad (1)$$

Recall that we wish to utilize knowledge from the ontology; one option is to fully exploit ontology structure. A CV term that is farther up in the ontology, towards the root, is typically a general concept and its presence in a cluster is less interesting to scientists. This is especially true if the cluster has CV terms at much greater depth. Our goal is to reduce the number of such general concepts that occur in the annotation signature. To do so, we define an extension of d_{tax} named d_{tax}^{str}; it will assign low values of ontological relatedness (similarity) to pairs of CV terms where at least one of the terms is a general concept in the ontology. Let *MaxDepth_Ontology* represent the greatest depth in the ontology.

$$d_{tax}^{str}(A,B) = d_{tax}(A,B) * (1 - pFactor(A,B)) \tag{2}$$

$$pFactor(A,B) = \frac{\max(correctedDepth(A), correctedDepth(B))}{MaxDepth\_Ontology}$$

$$correctedDepth(X) = MaxDepth\_Ontology - Depth(X)$$

Definition 1 (Cluster Density). *Given a labeled bipartite graph BG=($A_i \cup A_j$, WE) with nodes A_i and A_j and edges WE, a distance metric d, and a subset p of WE, the cluster density of p $cDensity(p) = \frac{\sum_{e \in p} 1-d(e)}{|p|}$.*

Definition 2 (The Annotation Signature Partition Problem). *Given a labeled bipartite graph BG=($A_i \cup A_j$, WE), a distance metric d, and a real number θ in the range [0.0:1.0]. For each $a \in A_i$ and $b \in A_j$, if 1-d(a,b) $> \theta$, then there is an edge $e = (a,b) \in$ WE. For each $e = (a,b) \in$ WE, label(e)= 1-d(a,b). The AnnSig Partition Problem identifies a (minimal) partition P of WE such that the aggregate cluster density P $AnnSig(P) = \frac{\sum_{p \in P}(cDensity(p))}{|P|}$ is maximal.*

AnnSigClustering is a greedy iterative algorithm to solve the *Annotation Signature* Partition Problem. *AnnSigClustering* adds an edge to a cluster following a greedy heuristic to create clusters that maximize the cluster density. *AnnSigClustering* assigns a score to an edge e in *WE* according to the number of edges whose adjacent terms are dissimilar to the terms of e, and that have been already assigned to a cluster. Then, edges are chosen in terms of this score (descendant order). Intuitively, selecting an edge with the maximum score, allows *AnnSigClustering* to place first the edges with more restrictions; this is one for which there is a smaller set of potential clusters. The selected edge is assigned to the cluster that maximized the cluster density function. Time complexity of *AnnSigClustering* is $O(|WE|^3)$. To illustrate the behavior of *AnnSigClustering*, lets consider the annotated graph in Figure 2. This graph can be partitioned into 2 groups of edges, e.g., one group includes the edges between `State Guarantor` on the left with two terms `STRIP` and `State Guarantor` on the right; also, the edge between `Municipal Debt Issuer` belongs to this group. The other group is comprised of edges between `Perpetual Maturity` on the left with two terms `Maturity Duration` and `Perpetual Maturity` on the right, as well as the edge between `Putable` and `Perpetual Maturity`. `Exchangeable` that is not ontologically related to any of the FIBO terms associated with the `Financial Contract A` (on the left). These two clusters were created because when each of the edges was assigned to the corresponding cluster, similarity values between the adjacent terms of all the edges in the clusters, were high enough to ensure that cluster density was maximized.

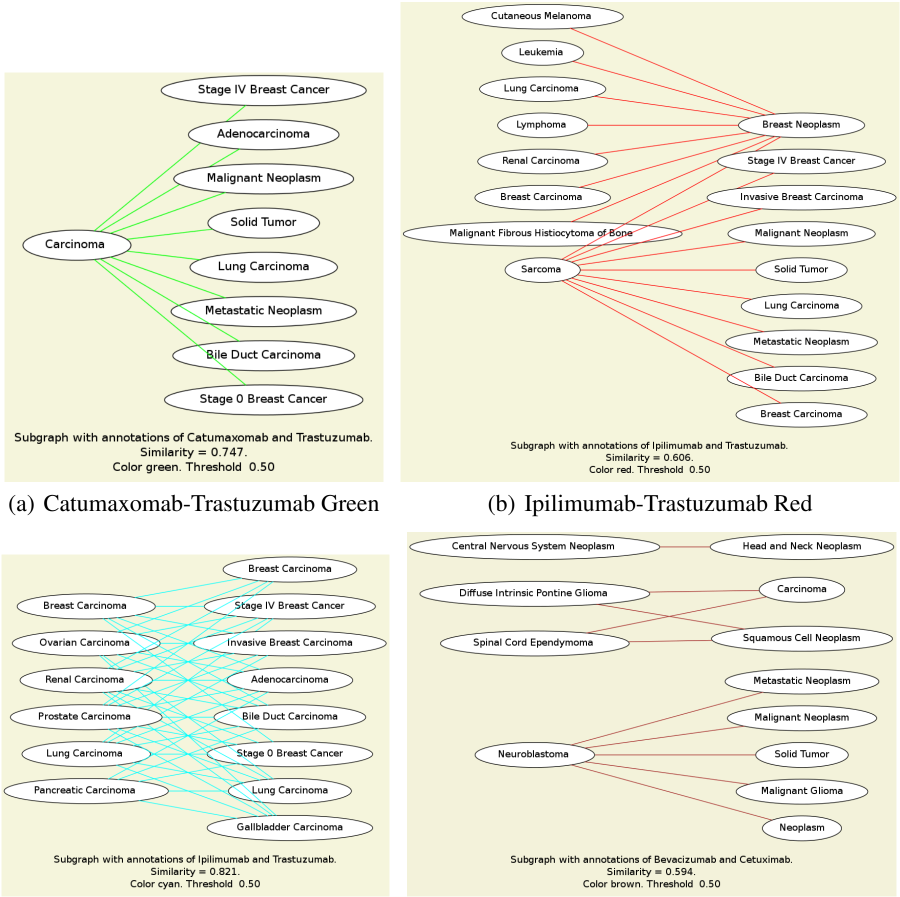
4 Evaluation

The goal of our evaluation is to validate if annotation signatures group together meaningful terms across shared annotations. Additionally, we evaluate the impact of the semantics encoded in the ontologies on the quality of the signature. We study two annotated datasets: *i*) Twelve drugs annotated with NCIt terms that correspond to the diseases associated with these drugs in clinical trials. *ii*) Twenty transporter genes from Arabidopsis thaliana annotated with GO terms. There is no prior *gold standard* solution(s) or ground truth for these two datasets that we can use to evaluate the quality of the annotation signature. Thus, we relied on a team of experts to analyze the annotation signatures. Annotated datasets are included in the supplementary material. All results are available via a Web portal [2].

4.1 Dataset and Evaluators

Drugs: Anti-neoplastic agents and monoclonal antibodies are two popular and independent intervention regimes that have been successfully applied to treat a large range of cancers. There are 12 drugs that fall within their intersection, and scientists are interested in studying the relationships between these drugs and the corresponding diseases. We consider a dataset of the following twelve drugs: `Alemtuzumab`, `Bevacizumab`, `Brentuximab vedotin`, `Cetuximab`, `Catumaxomab`, `Edrecolomab`, `Gemtuzumab`, `Ipilimumab`, `Ofatumumab`, `Panitumumab`, `Rituximab`, and `Trastuzumab`. The protocol to create the dataset is as follows: Each drug was used to retrieve a set of clinical trials in LinkedCT *circa* September 2011 (`linkedct.org`). Then each disease associated with each trial was linked to its corresponding term in the NCI Thesaurus version 12.05d; annotation was performed by NCIt experts. Our group of evaluators included two experts who develop databases and tools for the NCI Thesaurus and two bioinformatics researchers with expertise on the NCIt and other biomedical ontologies.

Genes: The vacuolar-type H+-ATPase are proton pumps associated with the `adenosine triphosphatase (ATP)` enzyme. The pump acidifies intracellular compartments and is essential for many processes, including co-transport, guard cell movement, development, and tolerance to environmental stress. Our collaborators in the Sze Lab at the University of Maryland have identified genes encoding subunits of `V-ATPase` in the `Arabidopsis thaliana genome`. The pump consists of subunits `A` through `H` of the `peripheral V1` complex, and subunits `a`, `c`, `c"` and `d` of the `Vo membrane` sector. The genes are named `AtVHA-n` where `n` represents the code for each subunit. Our dataset included the following twenty genes, `AtVHA-A`, `AtVHA-A1`, `AtVHA-A2`, `AtVHA-A3`, `AtVHA-B1`, `AtVHA-B2`, `AtVHA-B3`, `AtVHA-C`, `AtVHA-C1`, `AtVHA-C2`, `AtVHA-C3`, `AtVHA-C4`, `AtVHA-C5`, `AtVHA-D1`, `AtVHA-D2`, `AtVHA-E1`, `AtVHA-E2`, `AtVHA-F`, `AtVHA-c"1` and `AtVHA-c"2`. We obtained the GO annotations from the TAIR portal[3].



(a) Catumaxomab-Trastuzumab Green

(b) Ipilimumab-Trastuzumab Red

(c) Ipilimumab-Trastuzumab Cyan

(d) Bevacizumab-Cetuximab Brown

Fig. 3. Connectivity Patterns within Each Cluster for $\theta = 0.5$; (a) Catumaxomab-Trastuzumab Green; (b) Ipilimumab-Trastuzumab Red; (c) Ipilimumab-Trastuzumab Cyan; (d) Bevacizumab-Cetuximab Brown.

4.2 Connectivity Patterns within a cluster

The connectivity pattern within each cluster provides insight into the ontological relatedness of the diseases. In Figure 3(a) `Carcinoma` on the left is connected to 8 terms on the right. In Figure 3(b), `Sarcoma` on the left is connected to 9 drugs on the right. Similarly, `Breast Neoplasm` on the right is connected to eight diseases on the left. None of the other drugs has more than one incident edge. In contrast, in Figure 3(c), we see a much more general many-to-many connection pattern between the diseases on the left and right. Finally, Figure 3(d) shows a more complex connectivity pattern where the terms are ontologically related but they are placed within three disconnected graphs. The four terms `Diffuse Intrinsic Pontine Glioma`, `Spinal`

[2] dynbigraph.appspot.com
[3] http://www.arabidopsis.org/,April-May2013

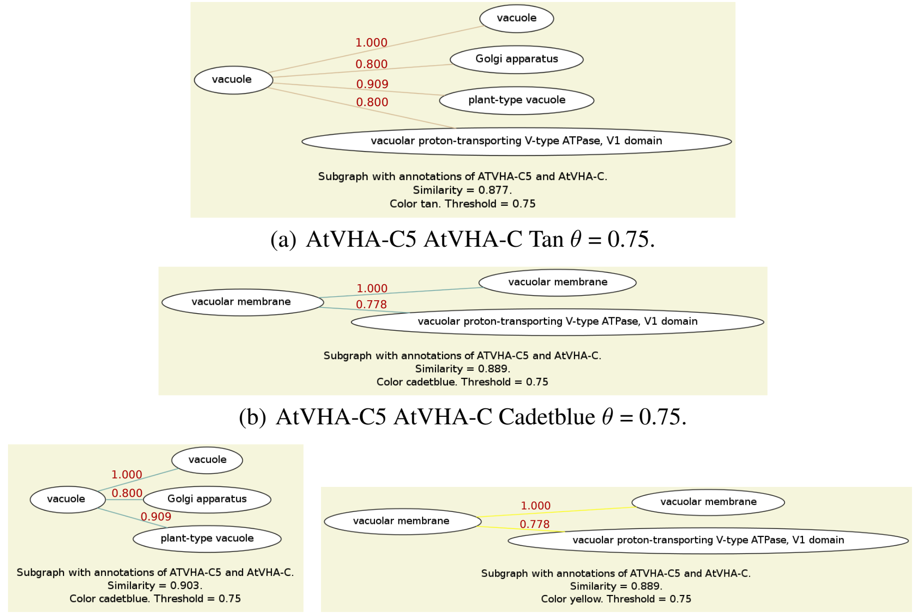
Cord Ependymoma, Carcinoma and Squamous Cell Neoplasm form the most well connected cluster. Comments from the evaluators noted that while groups such as Figure 3(a) that included generic terms such as Carcinoma were valid, they did not convey useful information. In contrast, groups in Figures 3(c) and (d), that had more specific terms and were more densely connected, had the potential to be more meaningful.

4.3 Utilizing Relationship Type Semantics

The goal of this evaluation is to determine the impact of the semantics of the ontology relationships on the annotation signatures. Figure 4 presents an example of exploiting relationship types using the GO ontology. There are five type of relationships captured in the GO ontology: *i)* is_a, *ii)* part_of, *iii)* regulates, *iv)* positively_regulates and *v)* negatively_regulates. Figures 4 (a) and (b) present two components of the gene signature for the genes AtVHA-C5 and AtVHA-C for threshold θ = 0.75. This is a scenario where d_{tax} (ontological relatedness) is computed using paths that consider *all* the GO relationship types. We observe that the term vacuolar proton-transporting V-type ATPase, V1 domain appears in both components of Figures 4(a) and (b). In contrast, Figures 4(c) and (d) present the two components when *only* is_a relationship types are considered. The value for d_{tax} between vacuole and vacuolar proton-transporting V-type ATPase, V1 domain decreases from 0.800 to 0.70. As a result, the term vacuolar proton-transporting V-type ATPase, V1 domain is only present in one component in Figure 4(d). This example illustrates multiple benefits from using ontological knowledge. First, redundancy in patterns is reduced. More important, the modified components represent more precise patterns of relationships between shared annotations and reflect additional semantic knowledge. A summary of this evaluation is described in Section 4.5.

4.4 Utilizing Ontology Structure

Recall that d_{tax}^{str} extended the taxonomic distance metric d_{tax} to consider ontology structure. Figure 5(a) illustrates an example cluster of the annotations for the pair Trastuzumab and Bevacizumab produced by d_{tax}; the threshold θ = 0.50. There are many shortcomings. First, it contains generic CV terms such as Adenocarcinoma and Carcinoma. Further, it is very large and many diverse and unrelated cancers are included. Figure 5(b) shows the result of applying the metric d_{tax}^{str} to exploit ontology structure. The large cluster was partitioned into smaller clusters. Many of the generic CV terms are no longer included and each smaller cluster includes more closely related CV terms. For example, one has a focus on breast cancer related terms, another has a focus on lung cancer, while a third combines terms related to pancreatic, renal and colorectal cancers. This example illustrates benefits from using ontological knowledge to eliminate generic terms from the annotation signatures. Redundancy in patterns is reduced, and the modified annotation signatures are comprised of relationships between more specific terms. Summarized results of the comparison between d_{tax} and d_{tax}^{str} for the dataset of the twelve drugs are presented in next section.


vacuole
1.000
0.800
0.909
0.800
vacuole
Golgi apparatus
plant-type vacuole
vacuolar proton-transporting V-type ATPase, V1 domain
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.877.
Color tan. Threshold = 0.75
vacuolar membrane
1.000
0.778
vacuolar membrane
vacuolar proton-transporting V-type ATPase, V1 domain
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.889.
Color cadetblue. Threshold = 0.75
vacuole
1.000
0.800
0.909
vacuole
Golgi apparatus
plant-type vacuole
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.903.
Color cadetblue. Threshold = 0.75
vacuolar membrane
1.000
0.778
vacuolar membrane
vacuolar proton-transporting V-type ATPase, V1 domain
Subgraph with annotations of ATVHA-C5 and AtVHA-C.
Similarity = 0.889.
Color yellow. Threshold = 0.75

(a) AtVHA-C5 AtVHA-C Tan $\theta = 0.75$.

(b) AtVHA-C5 AtVHA-C Cadetblue $\theta = 0.75$.

(c) AtVHA-C5 AtVHA-C Cadetblue $\theta = 0.75$ Only ISA Paths.

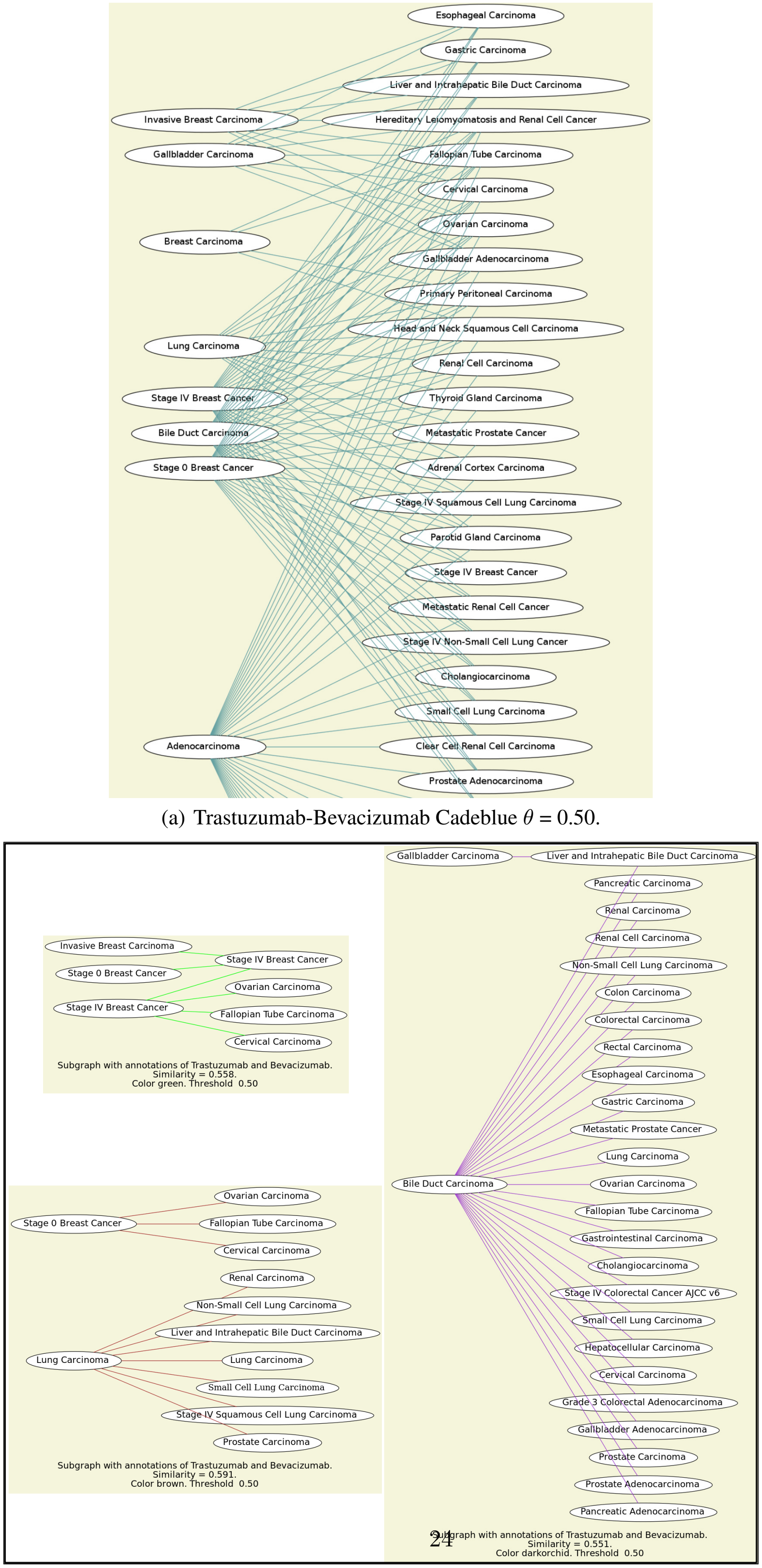
(d) AtVHA-C5 AtVHA-C yellow $\theta = 0.75$ Only ISA Paths.

Fig. 4. Enhancing Discovery Patterns with Semantics for $\theta = 0.75$; (a) and (b) Paths are computed using all five GO relationship types; (c) and (d) Paths are computed using only the `is_a` GO relationship type.

4.5 Summary Statistics

In this section we report on aggregated results of our evaluations. Table 1(a) provides a summary of the gene clustering when d_{tax} (ontological relatedness) is computed using all the GO relationship types and when only `IS_A` relationship types are considered. We compute the annotation signatures for pairwise comparisons of twenty genes; we report on minimum (MIN), maximum (MAX), and average (AVG) number of clusters in these annotation signatures. We consider two different values of threshold $\theta = 0.5$ and 0.75. As the threshold θ increases, the average of the number of clusters decreases. Further, when only `IS_A` relationship types are considered, the values of d_{tax} (ontological relatedness) are affected. The number of paths between two terms decreases, e.g., paths combining `positively_regulates` and `negatively_regulates` are not included in the bipartite graph BG. Additionally, d_{tax} values typically decrease and more edges are deleted from BG. Thus, as observed in Table 1(a), the average of number of clusters decreases. As noted earlier, these refinements also create more closely related clusters.

Table 1(b) provides summary statistics for annotation signatures computed using d_{tax} and d_{tax}^{str} over the pairwise comparisons of twelve diseases. We report on minimum (MIN), maximum (MAX), and average (AVG) number of clusters in these signatures; two values of threshold $\theta = 0.5$ and 0.75 are considered. Because d_{tax}^{str} penalizes generic CV terms, many edges are eliminated from BG. Further, the values of d_{tax}^{str} are lower than the values of d_{tax}. Thus, many of the large clusters computed with d_{tax} are partitioned into smaller clusters by d_{tax}^{str}. At the same time, the number of clusters de-


Esophageal Carcinoma
Gastric Carcinoma
Liver and Intrahepatic Bile Duct Carcinoma
Invasive Breast Carcinoma
Hereditary Leiomyomatosis and Renal Cell Cancer
Gallbladder Carcinoma
Fallopian Tube Carcinoma
Cervical Carcinoma
Ovarian Carcinoma
Breast Carcinoma
Gallbladder Adenocarcinoma
Primary Peritoneal Carcinoma
Head and Neck Squamous Cell Carcinoma
Lung Carcinoma
Renal Cell Carcinoma
Stage IV Breast Cancer
Thyroid Gland Carcinoma
Bile Duct Carcinoma
Metastatic Prostate Cancer
Stage 0 Breast Cancer
Adrenal Cortex Carcinoma
Stage IV Squamous Cell Lung Carcinoma
Parotid Gland Carcinoma
Stage IV Breast Cancer
Metastatic Renal Cell Cancer
Stage IV Non-Small Cell Lung Cancer
Cholangiocarcinoma
Small Cell Lung Carcinoma
Adenocarcinoma
Clear Cell Renal Cell Carcinoma
Prostate Adenocarcinoma

(a) Trastuzumab-Bevacizumab Cadeblue $\theta = 0.50$.

Invasive Breast Carcinoma
Stage IV Breast Cancer
Stage 0 Breast Cancer
Ovarian Carcinoma
Stage IV Breast Cancer
Fallopian Tube Carcinoma
Cervical Carcinoma
Subgraph with annotations of Trastuzumab and Bevacizumab.
Similarity = 0.558.
Color green. Threshold 0.50
Ovarian Carcinoma
Stage 0 Breast Cancer
Fallopian Tube Carcinoma
Cervical Carcinoma
Renal Carcinoma
Non-Small Cell Lung Carcinoma
Liver and Intrahepatic Bile Duct Carcinoma
Lung Carcinoma
Lung Carcinoma
Small Cell Lung Carcinoma
Stage IV Squamous Cell Lung Carcinoma
Prostate Carcinoma
Subgraph with annotations of Trastuzumab and Bevacizumab.
Similarity = 0.591.
Color brown. Threshold 0.50
Gallbladder Carcinoma
Liver and Intrahepatic Bile Duct Carcinoma
Pancreatic Carcinoma
Renal Carcinoma
Renal Cell Carcinoma
Non-Small Cell Lung Carcinoma
Colon Carcinoma
Colorectal Carcinoma
Rectal Carcinoma
Esophageal Carcinoma
Gastric Carcinoma
Metastatic Prostate Cancer
Lung Carcinoma
Bile Duct Carcinoma
Ovarian Carcinoma
Fallopian Tube Carcinoma
Gastrointestinal Carcinoma
Cholangiocarcinoma
Stage IV Colorectal Cancer AJCC v6
Small Cell Lung Carcinoma
Hepatocellular Carcinoma
Cervical Carcinoma
Grade 3 Colorectal Adenocarcinoma
Gallbladder Adenocarcinoma
Prostate Carcinoma
Prostate Adenocarcinoma
Pancreatic Adenocarcinoma
Subgraph with annotations of Trastuzumab and Bevacizumab.
Similarity = 0.551.
Color darkorchid. Threshold 0.50

(b) Trastuzumab-Bevacizumab $\theta = 0.50$ using d_{tax}^{str}

Fig. 5. Enhancing Signatures with Semantics for $\theta = 0.50$. (a) Signature of Trastuzumab-Bevacizumab $\theta = 0.50$; Similarity d_{tax}-Figure has been truncated for readability.; (b) Three clusters of Trastuzumab-Bevacizumab $\theta = 0.50$ when generic terms are penalized using d_{tax}^{str}.

Table 1. Cluster Distribution over the set of annotated drugs and genes. (a) Aggregate Cluster Distribution, all GO relations versus only is_a for AtVHA-n genes ; (b) Aggregate Cluster Distribution, effect of d_{tax} versus d_{tax}^{str} to eliminate relationships with generic terms. MIN, MAX, AVG correspond to the minimum, maximal and average numbers of clusters identified by *AnnSigClustering*, respectively.

(a) Aggregate Cluster Distribution AtVHA-n genes

	MIN	MAX	AVG
0.50	4.00	28.00	**11.96**
0.50 Only is_a	4.00	30.00	**11.75**
0.75	2.00	34.00	**10.99**
0.75 Only is_a	2.00	34.00	**10.45**

(b) Aggregate Cluster Distribution Diseases

	MIN	MAX	AVG
0.50 d_{tax}	1.00	46.00	**6.26**
0.50 d_{tax}^{str}	0.00	28.00	**3.38**
0.75 d_{tax}	0.00	37.00	**4.92**
0.75 d_{tax}^{str}	0.00	9.00	**0.80**

creases. All of these refinements lead to a smaller number of more closely related and meaningful clusters within the annotation signature.

5 Related Work

Graph data mining [5] covers a broad range of methods dealing with the identification of (sub)structures and patterns in graphs. Popular techniques include graph clustering, community detection and cliques. The problem of a 1-to-1 weighted maximal bipartite match has been applied to many problems, e.g, semantic equivalence between two sentences and measuring similarity between shapes for object recognition[1, 3, 11]. These approaches clearly show the benefits of solving a matching problem to identify similarity between terms or concepts. Our research advances prior research in that we consider the relatedness of sets of annotations and identify a many-to-many bipartite match.

A key element in finding patterns is identifying related concepts; we consider ontological relatedness. Similarity metrics (or distance metrics) can be used to measure relatedness; we briefly describe some of the existing metrics. The first class of metrics are string-similarity[4]; they compare the names or labels of the concepts using string comparison functions based on edit distances or other functions that compare strings. This includes the Levenstein distance and Jaro-Winkler [6]. The next are path-similarity metrics that compute relatedness based on the paths that connect the concepts within some appropriate graph. Nodes in the paths can be all of the same abstract types (e.g., PathSim [13]) or they can be heterogeneous (HeteSim [12]). Furthermore, topological-similarity metrics extend the concept of path-similarity and they look at relationships within an ontology or taxonomy that is itself designed to capture relationships (e.g., nan [7], d_{ps} [9] and d_{tax}[2]). We propose an approach that exploits ontological knowledge of scientific annotations to decide relatedness between entities of annotated datasets.

6 Conclusions and Future Work

We have defined the *Annotation Signature* Partitioning problem and the *AnnSigClustering* algorithm to develop the components of a signature based on shared annotations and ontological relatedness. We empirically studied the effectiveness of *AnnSigClustering* to identify potential meaningful signatures of annotated concepts. Further, we have analyzed the effects of considering knowledge encoded in the ontologies used to annotate Linked Data. Our results suggest that the grouping capability of our approach is enhanced whenever the type of relationships are considered as well as when relationships with generic terms are eliminated. Our initial project objective was to validate correctness and utility of components in a signature. Nevertheless, in the future, we will also address performance and scalability. Additionally, we plan to conduct a deeper evaluation study with our collaborators, and thus determine the potential discovery capability of the approach. Finally, we plan to apply our techniques to other domains, e.g., to identify signatures of electoral voters, relationships between financial contracts, and patterns of viral diseases.

References

1. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002.
2. J. Benik, C. Chang, L. Raschid, M. E. Vidal, G. Palma, and A. Thor. Finding cross genome patterns in annotation graphs. In *Proceedings of Data Integration in the Life Sciences (DILS)*, 2012.
3. S. Bhagwani, S. Satapathy, and H. Karnick. Semantic textual similarity using maximal weighted bipartite graph matching. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 579–585. Association for Computational Linguistics, 2012.
4. W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, pages 73–78, 2003.
5. D. J. Cook and L. B. Holder. *Mining graph data*. Wiley-Blackwell, 2007.
6. M. A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, pages 491–498, 1995.
7. B. McInnes, T. Pedersen, and S. Pakhomov. Umls-interface and umls-similarity : Open source software for measuring paths and semantic similarity. *Proceedings of the AMIA Symposium*, pages 431–435, 2009.
8. G. Palma, M.-E. Vidal, E. Haag, L. Raschid, and A. Thor. Measuring relatedness between scientific entities in annotation datasets. Technical report, University of Maryland. UMIACS Technical Report, 2013.
9. V. Pekar and S. Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *COLING*, 2002.
10. L. Raschid, G. Palma, M.-E. Vidal, and A. Thor. Exploration using signatures in annotation graph datasets. Technical report, University of Maryland. UMIACS Technical Report, 2013.
11. Y. Shavitt, E. Weinsberg, and U. Weinsberg. Estimating peer similarity using distance of shared files. In *International workshop on peer-to-peer systems (IPTPS)*, volume 104, 2010.
12. C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu. Relevance search in heterogeneous networks. In *EDBT*, pages 180–191, 2012.
13. Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.

Exploiting Semantics from Ontologies and Shared Annotations to Find Patterns in Annotated Linked Open Data (Supplementary Material)

Guillermo Palma[1], Maria-Esther Vidal[1], Louiqa Raschid[2], and Andreas Thor[3]

[1] Universidad Simón Bolívar, Venezuela
[2] University of Maryland, USA
[3] University of Leipzig, Germany
gpalma@ldc.usb.ve, mvidal@ldc.usb.ve, louiqa@umiacs.umd.edu, thor@informatik.uni-leipzig.de

1 NCIt annotations of the twelve drugs in the intersection of Anti-neoplastic agents and monoclonal antibodies

`Alemtuzumab:` C4376, C3149, C4337, C2985, C3208, C3247, C15194, C3243, C3242, C12981, C40022, C17600, C8851, C15342, C34375, C15265, C15289, C15261, C9300, C35069, C9385, C75570, C99382, C3468, C34383, C3211, C3092, C3196, C9438, C9357, C27134, C1681, C12415, C82650, C3063, C3167, C3161, C3163, C15271

`Bevacizumab:` C9270, C3773, C4337, C27977, C3739, C27971, C3242, C27806, C3088, C4436, C40022, C4822, C35018, C4908, C9063, C35468, C2955, C8563, C7431, C12945, C3158, C9166, C3400, C27962, C4917, C3099, C3552, C26874, C9477, C4910, C4911, C4912, C3809, C84457, C51302, C27970, C8767, C7927, C2907, C3414, C3262, C3261, C4033, C4049, C3796, C3814, C3815, C2919, C8946, C26782, C3813, C96963, C4326, C34447, C26766, C6791, C9384, C9385, C9382, C94764, C2916, C2910, C3270, C3861, C3043, C4863, C3867, C4878, C8411, C4872, C3382, C89999, C2929, C9039, C3568, C2926, C3209, C3208, C3364, C8925, C6959, C34794, C3207, C34982, C28194, C3875, C8524, C9145, C84391, C9306, C2956, C9305, C2852, C62332, C3353, C34538, C3350, C2953, C3513, C4013, C4012, C3194, C4815, C3161, C3163, C3044, C50837, C12341, C9112, C3224, C3995, C9118, C3850, C4656, C8294, C8566, C4005, C7511, C9325, C2039, C19151, C3058, C3059, C9448, C3117, C36263, C3234, C7510, C34863, C35064, C3538, C3108, C9292, C9293, C85218, C8516

`Brentuximab vedotin:` C3211, C3720, C9357

`Catumaxomab:` C4911, C4984, C2916, C3815, C4908, C2885, C4004

`Cetuximab:` C2929, C7927, C3995, C9039, C9270, C3998, C9118, C2926, C3850, C3262, C3261, C4043, C3200, C35850, C4025, C3792, C4758, C8543, C7558, C9238, C3871, C19151, C3058, C2953, C3291, C2956, C3077, C2955, C7431, C9305, C34447, C4822, C9382, C9383, C2916, C90016, C5105, C4349, C3099, C3513, C4024, C4910, C4911, C9292, C4878, C4855, C4872, C8516, C89999, C4978

`Edrecolomab:` C2955

Gemtuzumab: C3171

Ipilimumab: C3224, C9118, C3850, C3208, C3247, C8925, C3242, C4863, C7918, C4908, C7712, C8563, C7087, C9384, C3510, C3270, C9292, C4878, C3161, C9096, C4872, C8591

Ofatumumab: C7876, C8141, C7875, C7847, C3208, C8565, C7874, C8070, C3209, C3465, C41168, C8115, C3163, C2884, C3211, C4341, C8646, C7540

Panitumumab: C3995, C4025, C2926, C3850, C3261, C89794, C7771, C9039, C4863, C2955, C9305, C9384, C9382, C3513, C4013, C4910, C17837, C4978, C9296, C4878, C8101, C4872

Rituximab: C3432, C9178, C26883, C4337, C3247, C15194, C2889, C3242, C2884, C26808, C7539, C2952, C2953, C34481, C21882, C27961, C3098, C4341, C3720, C26925, C27006, C84934, C3121, C84939, C7400, C84389, C3305, C3149, C53529, C80307, C3898, C78797, C2983, C3200, C27153, C4967, C26784, C8073, C8070, C34845, C34909, C26760, C2912, C2910, C7402, C3270, C34995, C26912, C3063, C27576, C27578, C7540, C3387, C4376, C75545, C3209, C3208, C3201, C60989, C61283, C3471, C35424, C84417, C8851, C46089, C3071, C15265, C7264, C9301, C26744, C85170, C21912, C3212, C3604, C3211, C3446, C3444, C9305, C18011, C27146, C3167, C3161, C3163, C27351, C8504, C21926, C3056, C7192, C4981, C3037, C75570, C34383, C62221, C9357, C9293, C34416, C61277, C26323, C15430, C3108

Trastuzumab: C3641, C98358, C3995, C9270, C9245, C98275, C9292, C52166, C2910, C17756, C3261, C4878, C4872, C9305, C3844, C2852, C16239, C27814

2 GO annotations of twenty transporter genes from Arabidopsis thaliana

AtVHA-A1: GO:0016887, GO:0012510, GO:0015078, GO:0005737, GO:0005794, GO:0005768, GO:0005802, GO:0005773, GO:0070070, GO:0033177

AtVHA-A2: GO:0033177, GO:0016887, GO:0031669, GO:0009678, GO:0009507, GO:0005774, GO:0043181, GO:0016020, GO:0005737, GO:0005794, GO:0045735, GO:0000325, GO:0015986, GO:0070072, GO:0005773, GO:0005739, GO:0009705, GO:0032119

AtVHA-A3: GO:0016887, GO:0006816, GO:0007030, GO:0009651, GO:0007033, GO:0005886, GO:0009678, GO:0032119, GO:0009705, GO:0005794, GO:0045735, GO:0000325, GO:0016020, GO:0070072, GO:0016049, GO:0009507, GO:0043181, GO:0005737, GO:0005773, GO:0005774, GO:0000902, GO:0031669, GO:0009941, GO:0015986, GO:0048193

AtVHA-A: GO:0006833, GO:0009266, GO:0006816, GO:0016820, GO:0009651, GO:0007033, GO:0005886, GO:0046933, GO:0046961, GO:0007010, GO:0006007, GO:0005774, GO:0009941, GO:0000325, GO:0016020, GO:0015991, GO:0015992, GO:0006972, GO:0016049, GO:0009506, GO:0009507, GO:0006098, GO:0048046, GO:0005773, GO:0005739, GO:0006094, GO:0006096, GO:0000902, GO:0007030, GO:0005618, GO:0009555, GO:0046686, GO:0005794, GO:0002020, GO:0010498, GO:0005524, GO:0046034, GO:0048193

AtVHA: GO:0033180, GO:0016820, GO:0051693, GO:0046686, GO:0005524, GO:0005886, GO:0010255, GO:0046961, GO:0030835, GO:0033178, GO:0016020, GO:0015991, GO:0015992, GO:0046933, GO:0016469, GO:0009506, GO:0009507, GO:0005773, GO:0005774, GO:0051015, GO:0051017, GO:0005794, GO:0046034

AtVHA-B2: GO:0033180, GO:0016820, GO:0051693, GO:0009651, GO:0005524, GO:0005886, GO:0046961, GO:0030835, GO:0005774, GO:0033178, GO:0007010, GO:0016020, GO:0015991, GO:0015992, GO:0046933, GO:0016469, GO:0009941, GO:0006098, GO:0005773, GO:0006094, GO:0051015, GO:0051017, GO:0005794, GO:0010498, GO:0046034

AtVHA-B3: GO:0046933, GO:0046034, GO:0033178, GO:0033180, GO:0051015, GO:0009507, GO:0051017, GO:0016820, GO:0051693, GO:0005794, GO:0005524, GO:0005886, GO:0016469, GO:0046961, GO:0005773, GO:0015991, GO:0005774, GO:0030835, GO:0015992

AtVHA-C"1: GO:0009651, GO:0016887, GO:0006970, GO:0033177, GO:0046686, GO:0015078, GO:0006816, GO:0007030, GO:0033179, GO:0007033, GO:0005773, GO:0015991

AtVHA-C1: GO:0006007, GO:0016887, GO:0033177, GO:0009507, GO:0015078, GO:0033179, GO:0005773, GO:0015991, GO:0046961, GO:0015992

AtVHA-C"2: GO:0016887, GO:0015991, GO:0009507, GO:0015078, GO:0033179, GO:0005773, GO:0033177

AtVHA-C2: GO:0016887, GO:0033177, GO:0009507, GO:0015078, GO:0005774, GO:0033179, GO:0005773, GO:0015991, GO:0046961, GO:0000220

AtVHA-C3: GO:0016887, GO:0015991, GO:0009507, GO:0015078, GO:0033179, GO:0005773, GO:0033177

AtVHA-C4: GO:0016887, GO:0015991, GO:0009507, GO:0015078, GO:0033179, GO:0005886, GO:0005773, GO:0033177

ATVHA-C5.annt: GO:0006007, GO:0016887, GO:0048767, GO:0033177, GO:0009507, GO:0015078, GO:0006816, GO:0007030, GO:0033179, GO:0009651, GO:0005886, GO:0005773, GO:0015991, GO:0005774

AtVHA-C.annt: GO:0016051, GO:0006816, GO:0016820, GO:0009651, GO:0007033, GO:0005886, GO:0046961, GO:0006007, GO:0043255, GO:0006511, GO:0000325, GO:0009932, GO:0015991, GO:0048765, GO:0016049, GO:0080129, GO:0009507, GO:0051788, GO:0009826, GO:0005773, GO:0005774, GO:0000221, GO:0000902, GO:0007030, GO:0009809, GO:0005794, GO:0009853, GO:0030243, GO:0048193

AtVHA-D1.annt: GO:0033177, GO:0009506, GO:0015078, GO:0046961, GO:0033179, GO:0005794, GO:0005886, GO:0000325, GO:0005773, GO:0015991, GO:0005774, GO:0015992

AtVHA-D2.annt: GO:0033177, GO:0009506, GO:0015078, GO:0046961, GO:0033179, GO:0005794, GO:0005773, GO:0015991, GO:0005774, GO:0015992

AtVHA-E1.annt: GO:0033179, GO:0006661, GO:0015991, GO:0005773, GO:0015078

AtVHA-E2.annt: GO:0015991, GO:0006661, GO:0015078, GO:0006816, GO:0007030, GO:0009651, GO:0005773, GO:0048193, GO:0006623, GO:0016192, GO:0006944

AtVHA-F.annt: GO:0046933, GO:0005634, GO:0033178, GO:0033180, GO:0046961, GO:0005794, GO:0005886, GO:0005773, GO:0015991, GO:0005774

Capturing intent and rationale for Linked Science: design patterns as a resource for linking laboratory experiments

Cameron McLean[1], Mark Gahegan[1], and Fabiana Kubke[2]

[1]University of Auckland, Centre for eResearch & Department of Computer Science
{ca.mclean,m.gahegan}@auckland.ac.nz
[2]University of Auckland, School of Medical Sciences
f.kubke@auckland.ac.nz

Abstract. The notion of design patterns, after architect Christopher Alexander, provides a powerful way to capture and describe reusable *design* knowledge for complex domains. In this position paper we present the idea of design patterns for molecular biology experiments, and discuss how they may be utilized to support experimental design reuse, reproducibility, and a platform for linking experiments. Design patterns provide an alternate terminology and interpretive framework that can capture expert experience and intent that is critical yet missing from current representations of lab methods that utilize web ontologies and computational workflows. We outline an approach to making design patterns a first class entity in support of linked experiments on the web and provide a glimpse of potential applications of laboratory design pattern knowledge.

Keywords: Linked Science, design patterns, semantics, ontology, workflows.

1 Introduction

While there has been much focus on the description and linkage of scientific datasets using web ontology languages, there has been less attention on how to describe and then link the surrounding laboratory methods that are an important step in generating such data. The use of biomedical ontologies[1] to annotate laboratory methods descriptions is a much needed and necessary first step to integrating laboratory experiments, however, current ontologies alone cannot always provide sufficient knowledge to support all the human reasoning and situated understanding one may need to act with such knowledge [1]. Traditionally, ontology takes in its remit the specification of domain semantics and hierarchical decomposition, while workflow representations encode processes - each effectively supporting the “what” and “how” of experiments respectively. Yet reusing an experimental design requires understanding of intent and rationale in addition to merely procedural facets - especially where they are to be executed in heterogeneous, non-computational (wet-lab) environments. This position paper introduces the concept of design patterns for laboratory experiments which can act as both container and notation to admit design rationale in a linked science setting.

[1] e.g. `http://www.bioontology.org/`

[2] For examples of lab patterns and use cases see `http://goo.gl/D5RZsQ`

2 Design patterns for laboratory experiments

Design patterns were first introduced in the domain of architecture by Christopher Alexander [2], as a way of encapsulating experts' knowledge and externalizing it to enable the generalization and communication of design. As a container for knowledge, patterns are realized as structured documents centered on problems, solutions, and the invariant "forces" that exist in a specified context. Through the invariant "forces", patterns identify, name, and abstract common themes in good design solutions that are gained from experience [3]. The pragmatic nature of design patterns, and their focus on expressing experience (rather than just domain concepts as for ontologies, or processes as for workflows) provides an architecture that can facilitate the adaption and reuse of laboratory experimental designs. Patterns provide a shared vocabulary and extensionally defined examples of solutions to complex problems and relate them back to an explicit rationale of why they are good.

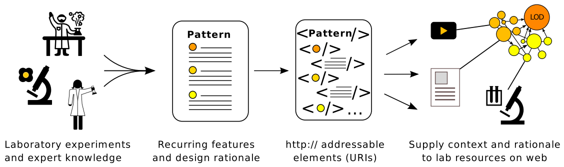

Pattern
<Pattern/>
LOD
Laboratory experiments and expert knowledge
Recurring features and design rationale
http:// addressable elements (URIs)
Supply context and rationale to lab resources on web

Fig1. Design patterns to capture laboratory knowledge.

Patterns not only give a set of concepts and vocabulary for a domain, but do so in a way that tells us what to do.[2] The problem/solution orientation of patterns gives us metaphorical dials to the domain, and tells us how to control and operate effectively with them. Unlike ontologies which aim to be capable of expressing any valid domain knowledge, patterns act more like recipes (yet more abstract and general than typical workflow representations) and give us a map[3] of the model space that tells us what parts we can vary, and what should remain invariant to achieve the desired outcome [4]. Considered as a form of knowledge management, we believe patterns offer additional advantages and add powerful metadata alongside traditional ontological and workflow approaches.

Other patterns exist in the context of the semantic web *e.g.* workflow[4] or ontology design patterns[5], which aim to provide usually *domain independent* constructs for normalizing and specifying knowledge modeling problems. In contrast, our notion of laboratory patterns as knowledge acquisition abstracts over laboratory procedures directly (*cf.* the modeling of them) to provide reusable design solutions for *scientific experiments* anchored in domain context - they supply us with domain concepts and relationships across diverse experiments gathered around a specified design intent.

[2] For examples of lab patterns and use cases see `http://goo.gl/D5RZsQ`

[3] A map analogy of patterns at `http://hillside.net/plop/2010/papers/kohls.pdf`

[4] `http://www.workflowpatterns.com/`

[5] `http://ontologydesignpatterns.org/wiki/Main_Page`

While we accentuate differences in representational approaches here, in reality we recognize the boundaries between ontologies, workflows, and design patterns are fuzzy as each tries to incorporate aspects of the other. For the purpose of discussion we make some general distinctions between the traditional forms of the three in Table 1 below.

Table 1. Some general distinctions between the traditional forms of workflows, ontologies, and design patterns for representing knowledge.

Property	**Workflows**	**Ontologies**	**Design Patterns**
Mode	Descriptive	Descriptive	Prescriptive/Instructive
Degree of formality	Formal	Formal	Typically not formal
Concepts defined	Procedurally	Intensionally	Extensionally
Focus	Specific procedure	General	Specific Problem
Utility	Replication and automation of processes	Model all feasible cases and compute inferences	Understand, adapt and reuse design solutions
Formal semantics	Yes	Yes	No
Models	Processes	Knowledge /Facts	Implementations

3 Making design patterns and their vocabularies web addressable entities.

The traditional form of design patterns are structured documents written in natural language. Thus, in order to transform them into a resource for linked science, a mechanism for publishing patterns and their vocabulary as defined web addressable entities is desired.

We view design patterns as data and ask how we may publish pattern knowledge following linked data practices. To begin we have developed a method for capturing pattern knowledge utilizing social methods adapted from other domains where design patterns are valid entities. The structured documents that result from "pattern mining" are collaboratively transferred to a semantic wiki based on the OntoWiki Application framework [5]. OntoWiki and its extensions enable the direct semantic content authoring of a knowledge base expressed in RDF, and provide for simple human and machine accessible interfaces for publishing linked data. Patterns entered by users become instances of a pattern model with defined syntax and semantics for pattern elements such as title, problem description, forces, context *etc*. The structure and URIs provided by the semantic wiki present an important first step in extending the form of design patterns from paper to a web based resource and supports the reuse of pattern content. Additionally, the wiki captures provenance, enables peer review, and serves attribution and credit for design pattern authors.

The challenges to this approach consist of specifying and refining the semantic formalization of pattern level concepts and their relations using RDFS, OWL and appropriate logics, and subsequently tailoring the OntoWiki Application framework. This work is non-trivial as patterns have complex, interrelated internal and external structures, and remains the current focus of our efforts.

4 A vision for the application of laboratory pattern knowledge

The annotation of lab procedural descriptions with vocabulary and context supplied by patterns is an obvious use of patterns as metadata. Currently, this coupling must be created manually due to the implicit nature of many pattern concepts, but the markup of existing documents or the authoring of future ones can be facilitated by adapting existing annotation tools such as Rightfield [6]. Laboratory methods and other data on the web indexed to patterns can provide an additional handle to browse, search, or filter methods at a granular level, across domains, and at the level of design intent – one which current semantics do not adequately provide

Patterns name invariant forces that exist in recurring lab scenarios and provide a valuable step towards the specification of minimal information reporting guidelines for diverse laboratory processes. Indeed, the need for "high-level abstractions of the components of experimental workflows" has been noted [7]. Furthermore, patterns resemble a wet-lab equivalent of abstract computational workflows described by [8].

Our vision is the creation of a laboratory pattern catalogue and web resource, providing scientists assistance in understanding, reusing, and adapting the diversity of published laboratory methods to their own needs. Principal in our approach is the publication of pattern content and vocabulary as linked data, such that it may be available for use anywhere on the semantic web.

We believe the problem/solution orientation of design patterns fits well with the cognitive processes of laboratory scientists when engaged with methods knowledge. In combination with workflows and ontologies, the pragmatic aspects of pattern knowledge can help provide a type of balancing – filling a representational gap in our methods descriptions somewhere between axiomized ontologies and workflows that can improve the epistemological adequacy of our scientific record.

5 References

1. Pike, W., and Gahegan, M. Beyond ontologies: Toward situated representations of scientific knowledge. *Int. Journal of Human-Computer Studies*, *65*(7), 674-688. (2007)
2. Alexander, C. The timeless way of building. New York: Oxford University Press. (1979)
3. May, D., & Taylor, P. Knowledge management with patterns. *Communications of the ACM*, *46*(7), 94–99. (2003)
4. Gamma, E., Helm, R., Johnson, R., & Vlissides, J.. Design patterns: Abstraction and reuse of object-oriented design. Springer Berlin Hiedelberg. (1993)
5. Heino, N., et al. Managing Web Content Using Linked Data Principles-Combining Semantic Structure with Dynamic Content Syndication. *Computer Software and Applications Conference, IEEE 35th Annual*. 245-250. (2011)
6. Wolstencroft, K., et al. RightField: embedding ontology annotation in spreadsheets. *Bioinformatics* 27(14), 2021-2022. (2011)
7. Taylor, C. F., et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature biotechnology* 26:8, 889-896. (2008)
8. Garijo, D., and Gil, Y. A new approach for publishing workflows: abstractions, standards, and linked data. *Proceedings of the 6th workshop on Workflows in support of large-scale science. ACM*. (2011)

A Checklist-Based Approach for Quality Assessment of Scientific Information

Jun Zhao[1], Graham Klyne[1], Matthew Gamble[2], and Carole Goble[2]

[1] Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK
jun.zhao, graham.klyne@zoo.ox.ac.uk
[2] Computer Science, University of Manchester, Manchester, M13 9PL, UK
m.gamble@cs.man.ac.uk, carole.goble@manchester.ac.uk

Abstract. The Semantic Web is becoming a major platform for disseminating and sharing scientific data and results. Quality of these information is a critical factor in selecting and reusing them. Existing quality assessment approaches in the Semantic Web largely focus on using general quality dimensions (accuracy, relevancy, *etc.*) to establish quality metrics. However, specific quality assessment tasks may not fit into these dimensions and scientists may find these dimensions too general for expressing their specific needs. Therefore, we present a checklist-based approach, which allows the expression of specific quality requirements, saving users from the constraints of the existing quality dimensions. We demonstrate our approach by two scenarios and share our lessons about different semantic web technologies that were tested during our implementation.

1 Introduction

Information quality assessment aims to provide an indication of the fitness of information. Most existing approaches perform the assessment by integrating assessment of a number of quality dimensions, such as accuracy, completeness, or believability. We argue that although such methodology provides a systematic framework to organise quality assessment, it leaves two outstanding issues: 1) the quality dimensions used are often too abstract and generic for expressing concrete quality requirements, and 2) constrained frameworks are often unable to address different uses a consumer may have for a common resource: data fit for one purpose might not be fit for another. Although quality dimensions are often specialised to support assessment requirements from a specific domain or task, *e.g.* as a formula to compute a quality value by using a certain set of information, such specialisation cannot always be flexible enough to support different quality needs that might arise from different tasks to be applied to the same information. For example, the set of information considered sufficient for supporting access to a linked data resource might not be enough for assessing its freshness. Users need a flexible way to express their different quality requirements according to the task at hand.

This paper addresses these issues by proposing a flexible and extensible data model to support explicit expression of quality requirements. We draw upon the idea of checklists, a well-established tool for ensuring safety, quality and consistency in complex operations, such as manufacturing or critical care [4, 10]. A checklist explicitly defines a list of requirements that must be fulfilled or assessed for a given task. In our checklist-based framework we provide an OWL ontology, the *Minim ontology*, to express quality requirements as RDF, and an assessment tool to evaluate the conformance of target data against a Minim checklist. We demonstrate Minim in practice by applying it to support two quality assessment scenarios: the quality of scientific data, and scholarly artefacts.

The contributions of this paper are: 1) presenting a flexible and extensible data model for explicitly expressing quality requirements according to users' assessment needs; and 2) providing a comparison of several state-of-the-art semantic web technologies in supporting quality assessment tasks, which are learnt from our practical experience. The Minim model presented in this work is an updated version of our previous work [14], which provide two new distinct features: 1) more explicit representation of individual quality requirement as a type of test; and 2) an extensible structure for users to add requirements or tests that are not defined in the model, in order to cope with new emerging requirements from their own domains.

2 Motivating Scenarios

In this section we present our motivating quality assessment scenarios from the scientific and scholarly publishing domains. The scenarios illustrate how our checklist framework can be used to support specific quality assessment tasks. Although these requirements could be fit into a conventional quality dimension, such as correctness or completeness, our approach saved the users from having to take the extra step of identifying the relevant quality dimensions, which is commonly required in an existing dimension-based methodology. Therefore, our scenarios highlight the advantage and convenience of being able to explicitly express the assessment requirements using our approach.

2.1 Quality assessment of scientific linked data

The volume of scientific data resources on the linked data web is rapidly expanding. However, their quality does not always stand up to scrutiny, an issue that is caused either by the linked data publication process or is intrinsic to the source data. **Scenario 1** shows how quality assessment can reveal a series of potential quality issues in a linked dataset that contains some basic metadata information about 7,572 chemical compounds. The dataset was used in a previous study [7] and it was created based on the InfoBox information of Wikipedia[3]. Because of

[3] `http://en.wikipedia.org/`

the potential incompleteness of the information available from these InfoBoxes, the resulting linked dataset can also have some potential quality issues. For example, according to domain-specific recommendations, each chemical compound must have one and only one IUPAC International Chemical Identifier (InChI). A quality requirement like this can be easily expressed using the cardinality test construct in our checklist model (see section 3) and an assessment can be automatically performed against all the chemical compounds in the dataset.

2.2 Quality assessment for scholarly communication

Scholarly communication refers to a principled method of making *scientific artefacts* available in order to support their more effective interpretation and reuse. These artefacts include data, methods or tools that were used to generate the findings reported, and providing sufficient information is key to achieving this goal. This is an ongoing quality challenge in scholarly communication that has not been fully addressed.

Scenario 2 uses quality assessment to help boost the effectiveness of scholarly communication in practice. `myExperiment.org` [5] is a popular workflow repository for sharing and releasing scientific workflows, which are important first-class scientific artefacts documenting protocols used to generate experimental results. Re-use of these workflows relies on adequate documentation to facilitate understanding and re-purposing.

A previous study analysed a representative selection workflows from myExperiment.org and drew out a minimal set of information that supports their re-execution [14]. This information, presented as a quality checklist, can be used to prompt workflow authors to provide better documentation about the workflows. This early intervention enhances the quality of scholarly communication.

2.3 Summary

No quality dimensions need be mentioned in the quality requirements of our scenarios. Instead, these requirements can be directly expressed using the constructs of our checklist data model, see sections 3 and 6. This provides a novel approach to quality assessment, in comparison to most of the existing work.

3 Approach

Our checklist-based assessment approach is based on two central pieces: 1) a container data model for encapsulating the RDF data/graph to be evaluated, and 2) the Minim data model, for representing quality requirements as a checklist.

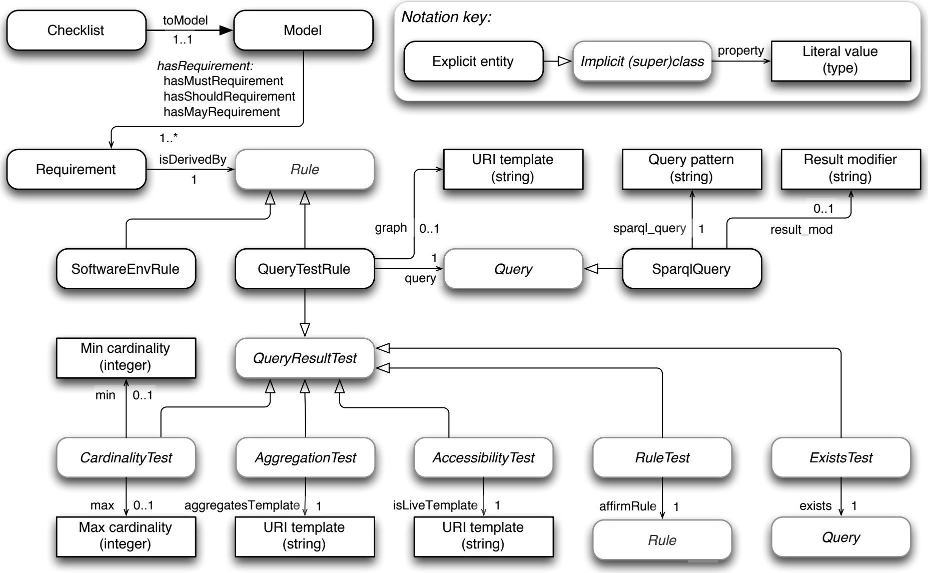
3.1 Research Object Model as a Container

We use an existing data model, namely the Research Object (RO) model [1], for our assessment. This provides a lightweight 'container' structure for encapsulating RDF and associated data. Annotation data contained within the RO constitutes the collection of RDF descriptions to be evaluated.

3.2 The Minim Model for Expressing Quality Requirements

A *checklist* provides an overall assessment of a dataset for some purpose. It consists of a number of individual *checklist items* which may address specific values within a dataset (typically at the level granularity accessible by a SPARQL query). Borrowing from IETF practice [4], individual items have a MUST, SHOULD or MAY requirement level. A dataset may be "fully compliant", "nominally compliant" or "minimally compliant" with a checklist if it satisfies all of its MAY, SHOULD or MUST items respectively.

Fig. 1. An overview of the Minim model schema.



Our Minim data model (see Figure 1) provides 4 core constructs to express a quality requirement:

- `minim:Checklist`[5], to associate a RO context, a target (the RO or a resource within the RO) and an assessment purpose (*e.g.* runnable workflow) with a `minim:Model` to be evaluated.
- `minim:Model`, to enumerate the requirements (checklist items) to be evaluated, with corresponding MUST, SHOULD or MAY requirement levels.
- `minim:Requirement`, which is a single requirement (checklist item) that is associated with a `minim:Rule` for evaluating whether or not it is satisfied or not satisfied.
- `minim:Rule`: There are several types of rules for performing different types of evaluation of the supplied data. Currently we have `minim:SoftwareEnvRule`, which tests to see if a particular piece of software is available in the current execution environment, and `minim:QueryTestRule`, which uses a query-based approach to assess the fitness of a target.

[4] `http://tools.ietf.org/html/rfc2119`

[5] The namespace of `minim` is `purl.org/minim/minim#`.

The following script, expressed using Turtle format, defines an example Minim checklist, which is to be used to assess each chemical compound must have exactly one InChI number. The checklist has one requirement that *must* be satisfied (line 9), i.e.,`:InChI`. The test of this rule is expressed by a SPARQL query (lines 19-20), which searches for the InChI identifier of a compound. The cardinality rule (lines 22-23) specifies that there must be exactly 1 matching query result associated with an evaluated compound.

```
 1 :runnable_workflow a minim:Checklist ;
 2   minim:forTargetTemplate "{+targetro}" ;
 3   minim:forPurpose "complete" ;
 4   minim:toModel :minim_model ;
 5   rdfs:comment """ Checklist to be satisfied if
 6          the chemical description is adequate.""" .
 7
 8 :minim_model a minim:Model ;
 9         minim:hasMustRequirement : InChI .
10
11 : InChI a minim:Requirement ;
12   rdfs:comment """Ensures exactly one chembox:StdInChI value
13     is defined on the target resource, and that its value is
14     a string literal.""" ;
15   minim:isDerivedBy [
16         minim:query
17                 [ a minim:SparqlQuery ;
18           minim:sparql_query
19            """?targetres chembox:StdInChI ?value .
20                 FILTER ( datatype(?value) = xsd:string ) """ ;
21                 ] ;
22                 minim:min 1 ;
23                 minim:max 1;
24                 minim:showpass "InChI identifier is present" ;
25                 minim:showfail "No InChI identifier is present" ;
26         ] .
```

In the current checklist implementation the `minim:QueryTestRule` is used to handle most of the checklist requirements we encounter. It can be associated with two elements: a query pattern (`minim:Query`) (lines 16-26), which is evaluated against the RDF data from the RO, and an optional external resource, which contains additional RDF statements that may be needed to complete the assessment. Every `minim:QueryTestRule` incorporates a `minim:QueryResultTest`, which takes the query result (which in our current case, a SPARQL query result) and returns a True (pass) or False (fail) result according to the type of test performed. Currently our Minim model defines 5 types of tests.

- `minim:CardinalityTest`, evaluates the minimum and/or maximum number of distinct matches in the query result against the declared conditions.
- `minim:AccessibilityTest`, evaluates whether a target resource indicated by the query result is accessible, by for example performing an HTTP HEAD request to the resource URI.
- `minim:AggregationTest`, tests the presence of resources in an RO that is used as the input to our assessment.
- `minim:RuleTest`, defines the additional rules to be applied to the assessment results returned from the evaluation of another `minim:QueryTestRule`. In this way, we can avoid writing too big rules and combine different types of rules, for example a query test rule with a liveness test rule.

- `minim:ExistsTest`, which can be used as a shortcut for a structure that combines a `minim:RuleTest` and `minim:CardinalityTest` to evaluate the existence of a particular resource in the evaluated data.

The Minim model is a refactor of our previous work [14], which addressed quality needs for enhancing scholarly communication (such as scenario 2). It has been extended by 1) explicitly defining an expandable set of test types; and 2) providing extension points allowing definitions of new assessment rules, assessment tests, and types of queries used to perform query-based tests (see *Rule*, *Query* and *QueryResultTest* in Figure 1).

Clearly, not every measure of quality can be evaluated automatically. For example, establishing correctness of stated facts may require independent validation [13]. Our approach allows direct tests to be combined with such independent validation or review, the latter of which may be simply expressed as quality metadata about the target dataset. A systematic assessment of how our checklist-based approach can support most of the existing known quality dimensions is a key part of our future work. Our focus on extensibility allows new automatic assessments to be introduced in a principled fashion. Examples of checklists that combine automatic evaluation with manual review may be found in our GitHub repository [6].

4 Implementation: The Minim Checklist Framework

The checklist framework is implemented in Python as both a command-line tool, `ro-manager`, and a RESTful service[7][8]. Source code is in GitHub[9].

As shown in Figure 2, the evaluation framework takes four inputs: a Research Object (RO) that containing a set of RDF annotations, a Minim file, a purpose indication, and an optional target resource URI (if not specified, the RO itself is the target). The framework uses a checklist from the Minim file selected by the purpose and target, applying each of the assessment tasks described by each checklist item to the RDF graph presented by the Research Object.

We chose SPARQL to express the `QueryTestRule`s within a Minim checklist, as SPARQL is a widely available standard for querying and accessing RDF data. Our comparison with other semantic web technology choices is presented in Section 6.

The assessment result contains quite extensive content in the form of an RDF graph. For web applications using these results, our implementation provides two additional services that return JSON or HTML checklist results that facilitate presentation of a more user-friendly “traffic-light”display, with “green ticks” for satisfied requirements, and “red crosses” and “yellow crosses” meaning failure of a MUST and SHOULD requirement respectively.

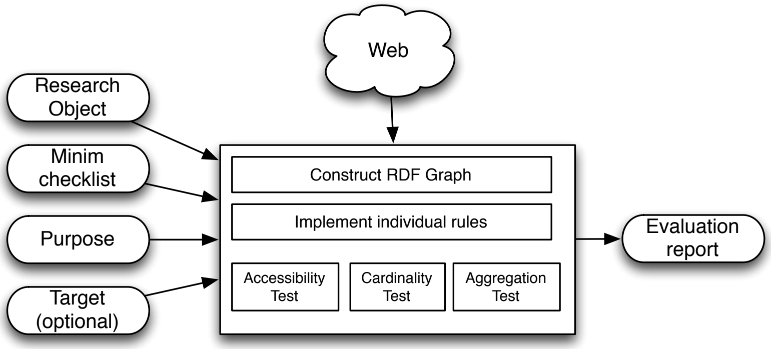
[6] `https://github.com/wf4ever/ro-catalogue/tree/master/minim`

[7] `http://purl.org/minim/checklist-service`

[8] Example REST service use is at `https://github.com/wf4ever/ro-catalogue/blob/master/minim/REST-invoke-checklist.sh`

[9] `https://github.com/wf4ever/ro-manager/`

Fig. 2. An outline of the checklist evaluation implementation



5 Quality Assessment in Action

In this section we show how the two motivating scenarios can be supported by our checklist tool. All the resources used for these case studies can be accessed in our Github repository[10]. Our exercise shows that our model and tool can sufficiently support assessment tasks from diverse domains, and at the same time, enable an explicit representation of the quality requirements from these tasks, which themselves can be valuable asset to a community.

5.1 Assess quality of scientific data using community checklist

In the first practical assessment we show how our checklist tool can be used to express existing community checklists from scientific domains in order to identify any potential quality issues of a scientific linked dataset. This actually reproduces the assessment by the previous MIM study [7] in our first motivating scenario. We reuse the chemical compound linked data and the checklist requirements defined in that study.

In that study 11 quality requirements were defined, based on a guideline from the chemistry domain. We analysed the tests required by each requirement[11] and categorised them into 3 different types: existence of information, type of information present, and cardinality of values provided. Our Minim model can be used to express these types of test, and the complete Mimim representation of these requirements is in our Github repository. We applied this checklist to 100 (limited by a performance constraint of the RO access mechanism used, currently being addressed) of the total 7,572 chemical compounds used in [7] and our checklist tool was able to reproduce exactly the same assessment result

[10] `http://purl.org/minim/in-use-submission/`

[11] `https://github.com/wf4ever/ro-catalogue/blob/master/v0.1/minim-evaluation/checklist-item-survey.md`

as the MIM checklist tool. Whilst we see this limited assessment as sufficient to demonstrate that we can reproduce the results of the MIM checklist, future work (discussed in Section 8) will include a full validation for completeness.

5.2 Assess quality of scholarly communication research objects for specific purpose

In our second case study we apply our checklist tool to a set of scientific workflows from the myExperiment.org repository. These workflows commonly rely on a third-party bioinformatics Web service provided by a research organisation in Japan[12]. At the end of year 2012, they announced that these services which were available as WSDL service would be upgraded to RESTful services and the WSDL service endpoints would no longer be supported, leading to failure of dependent workflows. Although it is impossible for them to be executable after the service upgrade, our assessment can enhance the quality of documentations about these workflows so that they can at least be understandable, repairable, and verifiable in the future.

Therefore, we designed a specific checklist, based on our previous analysis of causes to workflow quality issues [14]. In the checklist we define a list of requirements to be assessed, including: the presence of all input data; the presence of the workflow definition file; the presence of provenance logs of previous runs; and the accessibility of all the Web services used in a workflow.

22 workflows from myExperiment.org were applicable to our test. Our assessment managed to ensure that all the required information was associated with each workflow (see the full assessment result in our Github repository). After the service update took place, our checklist tool was able to successfully detect quality degradation for all the workflows and highlight explicitly the set of problematic services which caused the workflow no longer executable (see an example assessment result[13]). The assessment can be reproduced using resources in our Github repository.

6 Discussions

As an approach that is substantially based on semantic web technologies, the goals and features of our checklist-based framework can be seen to overlap with some major semantic web technologies like the Web Ontology Language (OWL) [14] and SPIN[15], which have been considered in our design process. However, our focus was to provide a higher level data model, which can more directly reflect quality requirements from users or specific scenarios. Although these semantic web technologies can be complementary to our approach, they cannot in isolation (fully) support all the quality assessment requirements identified from our scenarios.

[12] http://www.genome.jp/kegg/

[13] http://tinyurl.com/btxdlmv - this is a live service link

[14] http://www.w3.org/TR/owl2-overview/

[15] http://spinrdf.org

6.1 Comparison with an OWL-based Approach

OWL ontologies support the description of classes that detail the features necessary for an individual data item to be a member of that class. These class descriptions are analogous to the description of requirements in our checklist. OWL also has an RDF serialisation and extends RDF semantics[16] to operate over RDF data. We can express our InChI requirement in OWL as follows:

```
Class: InChI
   SubClassOf: chembox:StdInChI some :InChIValue .
```

However, the current OWL 2 RDF semantics contain two features that are incompatible with our quality checking scenario:

- The Open World Assumption (OWA). If an InChI were to be defined without a corresponding InChIValue, this would not be highlighted as an error by an OWL reasoner. Instead the OWA results in the inference that there exists an InChIValue, but that it is currently unknown. This directly conflicts with our need for an existence check.
- No Unique Names Assumption. We can extend the above requirement to include a cardinality restriction to say that there must be one and only one InChIValue. The presence of two different InChI values would not however raise an error. Instead the assumption would be made that the two InChIValues are in fact the same. This directly conflicts with our need for cardinality checks in a quality assessment scenario.

An alternative to the traditional OWL 2 Semantics are Integrity Constraint Semantics (ICs)[17]. ICs are a semantics for OWL that employ a Closed World Assumption as well as a form of the Unique Names assumption. These semantics therefore allow the use of OWL classes to be interpreted as integrity constraints. The Stardog database[18] currently provides an implementation of OWL with ICs.

One practical implementation of ICs is achieved by transforming the OWL classes to SPARQL queries. Each axiom in an OWL IC Ontology is transformed into a corresponding SPARQL query. This ability to realise ICs as SPARQL queries implies that by supporting a SPARQL based approach for requirement description, Minim achieves at least some of the expressiveness as an approach based upon OWL ICs. However, a purely OWL ICs based approach presents a number of restrictions with respect to what can be expressed in our requirements:

- Expression of different requirement levels such as MUST, SHOULD, and MAY. OWL IC semantics are primarily concerned with binary *satisfiability*, where we capture more nuanced levels of satisfaction. We believe would be more difficult to create checklists in OWL that capture these.

[16] http://www.w3.org/TR/rdf-mt/#MonSemExt

[17] http://stardog.com/docs/sdp/icv-specification.html

[18] http://www.stardog.com/

- Flexibility and extensibility to perform broader resource accessibility and software environment tests that can be supported by our Minim tool. For example verifying the web-accessibility of workflow input files lies outside the expressive scope of OWL (though might conceivably be handled through the introduction of new primitive classes and OWL resoner extensions).
- Expressing rules that validate data literal values. This has previously been highlighted as a restriction of an OWL based approach to data validation in the life sciences [3].

6.2 Comparison with a SPIN-based Approach

SPIN iprovides a query-based modelling language to express rules and logical constraints over RDF data. It is used by the previously discussed MIM checklist-based assessment framework.

The property of `spin:constraint` can support a set of features in common with our Minim tool. `spin:constraint` can be associated with an `rdfs:Class`, *e.g.* `chembox:InCHI`, and defines the constraints that instances of the class should comply with. The constraints can be expressed using SPARQL ASK or CONSTRUCT queries that are expressed using SPIN syntax in RDF. This structure can be used to support most of our query-based tests, apart from the accessibility tests. Additionally, `spin:Template`, which provides a meta-modelling function to group SPARQL queries so that they can be reused, is very similar to the role of `minim:Rule` in our model. However, at the time of writing, SPIN was not yet established as a standard and implementations of SPIN engines were limited. A purely SPIN-based approach also shares the first two restrictions as an OWL ICs based approach, as analysed above.

6.3 Summary

OWL, OWL ICs, and SPIN are clearly complementary to our Minim model approach. Although they cannot be directly used to support expressing quality assessment requirements, they can complement our SPARQL-based implementation of the checklist tool. SPARQL was chosen for our tool implementation because it is a more established standard for querying RDF data, with a number of known implementations. Combined with our Minim model, SPARQL can support all the expression of constraints and most of the inference functions as SPIN. However, our Minim model can also be extended and implemented using these alternative technologies. The `minim:Query` class is one extension point for supporting SPIN-like queries, and `minim:Rule` can be extended to define other than query-based test rules.

7 Related Work

Zaveri et al. [13] provides a timely and extensive survey on quality assessment of linked data. The survey is mainly organised by quality dimensions rather than the actual methodologies used by the reviewed works. Of the 21 works included in the review, a larger portion of them are based on specific algorithms, such as the trust evaluation by Golbeck [8] , or use a dimension-driven approach, such as Bizer et al [2], or take a purpose-built approach to provide solutions to a specific problem in a specific application scenario, such as Guéret et al. [9]. 3 of the works take an approach more closely related to ours by supporting an explicit expression of quality requirements. However, the quality schema provided by Sieve [12] is rather simple, mainly targeted to express the configuration parameters and the functions to be used for the assessment; and the quality ontologies proposed by SemRef [11] and SWIQA [6] are based on a series of quality dimensions.

8 Conclusions and Future Work

Quality assessment is a paramount issue in supporting the successful re-use of Scientific Linked Data. Not being able to express specific quality assessment requirements according to the needs from specific assessment tasks has been a bottleneck to the quality enhancement of linked data resources. To fill in this critical gap, we propose a checklist-based approach that allows explicit expression of quality requirements that can directly reflect users' needs from their concrete quality assessment tasks, and at the same provides flexible extensibility to cope with new needs. We show how our approach can support two exemplar case studies from scientific domains. We learnt valuable lessons about how various state-of-the-art semantic web technologies could support our concrete use in practice. The very lightweight SPARQL-based implementation has shown great promise in supporting these practical needs.

Our next steps will focus on the extensibility of the tool architecture, by exploring the possibility of a plug-in framework to enable plugging-in of third-party services. We are also prototyping a user interface tool to facilitate the creation of Minim checklists. Finally we are planning a systematic mapping between the existing quality dimensions and the constructs available in our checklist data model, to extend the function evaluation of our model.

References

1. Khalid Belhajjame, Oscar Corcho, Daniel Garijo, Jun Zhao, and et al. Workflow-centric research objects: First class citizens in scholarly discourse. In *Proceeding of SePublica2012*, pages 1–12, 2012.
2. Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(1):1–10, 2009.
3. Jerven Bolleman, Alain Gateau, Sebastien Gehant, and Nicole Redaschi. Provenance and evidence in uniprotkb. In *Proceedings of the 3rd International Workshop on Semantic Web Applications and Tools for the Life Sciences*, Berlin, Germany, 2010.
4. Sangyoon Chin, Kyungrai Kim, and Yea-Sang Kim. A process-based quality management information system. *Automation in Construction*, 13(2):241–259, 2004.

5. D. De Roure, C. Goble, and R. Stevens. The design and realisation of the myexperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25:561–567, 2009.
6. Christian Fürber and Martin Hepp. Swiqa–a semantic web information quality assessment framework. In *Proceedings of European Conference on Information Systems*, 2011.
7. Matthew Gamble, Carole Goble, Graham Klyne, and Jun Zhao. Mim: A minimum information model vocabulary and framework for scientific linked data. In *E-Science (e-Science), 2012 IEEE 8th International Conference on*, pages 1–8. IEEE, 2012.
8. Jennifer Golbeck and Aaron Mannes. Using trust and provenance for content filtering on the semantic web. In *Proceedings of the Models of Trust for the Web Workshop*, 2006.
9. Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In *Proceedings of the European Semantic Web Conference*, pages 87–102. Springer, 2012.
10. Brigette M. Hales and Peter J. Pronovost. The checklist–a tool for error management and performance improvement. *Journal of critical care*, 21(3):231–235, 2006.
11. Yuangui Lei, Victoria Uren, and Enrico Motta. A framework for evaluating semantic metadata. In *Proceedings of the 4th international conference on Knowledge capture*, pages 135–142. ACM, 2007.
12. Pablo N Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.
13. Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment methodologies for linked open data. *Semantic Web Journal*. submitted on 12/14/2012.
14. Jun Zhao, Jose Manuel Gomez-Perez, Khalid Belhajjame, and et al. Why workflows break-understanding and combating decay in taverna workflows. In *IEEE eScience*, pages 1–8, 2012.

A semantic Lab Notebook – Report on a Use Case Modelling an Experiment of a Microwave-based Quarantine Method

Nico Adams, Armin Haller, Alexander Krumpholz, Kerry Taylor

CSIRO, firstname.lastname@csiro.au

Abstract. A recent trend in a number of academic disciplines is the publication of results of experiments together with the scientific article for a better reproducibility of the published experiments and algorithms. Semantic Web technologies have the potential to aid scientists in the publishing, sharing and interlinking of this data and also in helping other scientists in the understanding of the data and the interpretation of the results of an experiment. In this paper we report on a use case on how to publish the data captured in a scientific experiment that has been conducted in the CSIRO Animal, Food and Health Sciences division as a set of ontologies and how to access this data through a set of RESTful semantic Web services. These services showcase how computational tasks that cannot be represented in the ontology can be implemented as lightweight semantic Web services to document and verify the results of an experiment. Together, the ontologies, the experimental data and the computational services constitute the elements needed for a semantically enabled lab notebook, facilitating research studies over multiple experiments, while reducing complexity and error rates.

1 Introduction

Research into different types of thermal treatments as quarantine methods against codling moth in a variety of fruit has gained much interest in recent years due to the uncertain future of chemical fumigation of food and the public concern over residues in treated products [25]. One such new thermal treatment method is currently under development in the CSIRO Animal, Food and Health Sciences division. The method proposes to use a combination of thermal treatment with microwave treatment of fruit to inactivate the fruit fly larvae from growth in different types of fruit. The experiment conducted by domain scientists in CSIRO is using a custom-built microwave-based heat treatment system that is tested on a number of different fruit for the inactivation of an induced codling moth infestation in these fruit.

Apart from the main goal of the research in establishing the effectiveness of the microwave-based heat treatment process for the inactivation of codling moth, a secondary goal of a transformational capability platform project was to showcase how the experimental data can be modelled in semantic Web languages and how the resulting ontological models can be used to support computational analysis of the experimental data. To achieve the latter, a group of ontologists and software engineers have accompanied the domain scientists during the experiment and defined models to capture the experimental data semantically. We present our methodology of how to publish the results of a scientific experiment as a set of ontologies. Further, we develop a set of services that showcase how computational tasks that cannot be represented in the ontologies themselves can be modelled and implemented as lightweight semantic Web services.

The remainder of this paper is structured as follows. In Sect. 2 we briefly describe the microwave-based heat treatment process that constitutes our use case. In Sect. 3 we describe the ontologies that are needed to capture this use case. In Sect. 4 we describe the information services that we have built on top of the ontological data representing the knowledge gathered in the use case experiment. We discuss some related work in Sect. 5, before we conclude in Sect. 6.

2 Microwave-based heat treatment

Our use case has been provided by scientists in the CSIRO Animal, Food and Health Sciences (CAFHS) who are developing a microwave-based heat treatment system for the treatment of fruit for the purposes of removing fruit fly larvae infestations. Briefly, for the experiment a purpose-built microwave tunnel system was built. The microwave unit also incorporates an auxiliary hot air system comprised of a heater and a fan which is also attached to the microwave system.

For the experiment that we modelled ontologically, newly harvested organically grown Mutsu and Granny Smith apples were used for treatment in this system. The apples were uniformly infested with fruit flies by making 50 pin holes on each apple at the stem end and then placed inside cages containing fruit flies.

The microwave treatment was applied by placing the apples on a small plastic stand with four protruding rods and sent through the microwave tunnel for approximately 54 min, which was preheated to $63 - 65°C$. A variable speed conveyor belt moves the fruit through the microwave tunnel where they undergo heating by microwaving to destroy fruit fly larvae and eggs. The temperature of the fruits at different points (top, flesh, core, bottom) were measured during the experiments with a fibre optic conditioner at 1 second intervals.

The goal of the experiment was to determine the optimal configuration of the tunnel temperature, the microwaving intensity and the time of treatment in each of the stages of the treatment process to obtain 100% mortality of the fruit fly larvae and eggs that is comparable to traditional thermal treatment methods. This can be done by calculating the cumulative thermal effect for a given treatment, based on kinetic data for the thermal mortality of target insects and for product quality losses. Given a time-temperature history of $T(t)$, the cumulative thermal mortality of the microwave-based heat treatment can be calculated to an equivalent length of time in minutes, M_{52}, at a reference temperature T^{ref} of $52°C$ by using the following relationship:

$$M_{52} = \int_0^t 10^{\frac{T(t)-52°C}{z}} dt$$

where M_{52} is the equivalent time at a target temperature of $52°C$, $T(t)$ is the transient temperature profile measured by the fibre optic system, t is the time and z is the temperature change (in $°C$) required to change the value of insect mortality (lethality) by a factor of 10.

3 A semantic lab notebook

To unambiguously record the data (e.g. temperature measurements, applied power, belt speed etc.) captured in the experiment and to allow a computational analysis of the experimental data, we first need a conceptual stratification of the experiment and a common understanding of the objects and processes that are used in the experiment. In the context of this project, a fruit treatment process use-case acts as an exemplar to build up a demonstrator of a semantics driven lab notebook.

When describing experiments in electronic lab notebooks the terms used are often ambiguous. In our use case experiment, for example, the term “fruit” if used in some electronic record, may be ambiguous depending on the experiment run, as there were multiple experiments conducted, not only with Mutsu and Granny Smith apples, but also with mangoes and avocados. Even more, many of these terms exhibit polysemy: avocado, in common usage - and therefore also when used as a metadata term - may, without further specification, refer to either the “avocado fruit” or the “avocado tree”. Such distinctions are important in that they (a) determine the scope of what we can talk about in our information systems and (b) also specify - at least to a degree what sort of data is to be collected. Completely defining the meaning of something allows the specification of the relationships between entities: an “avocado fruit”, for example, is part of an “avocado tree” (at least until it has been harvested). Such a disambiguation will then, for example, allow us to talk about properties of a specific fruit (e.g. volume, firmness) and its history and provenance (this “avocado fruit” was part of an “avocado tree” which was located in a “field” which is described by “geo-coordinates” X and Y etc.). Without disambiguating the polysemous term “avocado” it would have been impossible to represent information about an avocado in such terms.

The precise definition of objects and their relationships can also help to overcome the stratification in conceptual models of the treatment system, i.e. its factory: an apple has a digital representation denoting the apple in an information system, i.e. in the electronic lab notebook – for the purposes of this description we will call it a “digital apple”. The “digital apple” is described by some “apple description”, which, in turn is a kind of “information content entity”. An “information content entity”, in turn, may be an input into a “model”, for example, a model describing the relationship between the apple volume and the required heating intensity to reach a certain core temperature in the apple.

In summary, a semantic lab notebook is an exercise in object management. For the purposes of the rest of this discussion, the term “object” denotes any entity that can be named or addressed. Objects may therefore be physical objects as well as data objects, computational service objects etc..

We have chosen to model these objects with the languages developed in the technology stack of the semantic Web [2]. With RDF(s) and OWL, the central components of the semantic Web stack, it is possible to attach a formal, i.e. “computable” representation of a conceptualisation of the nature of the object to the object itself. Such statements can then be evaluated by reasoners which can draw inferences over the knowledge provided. We have developed a set of ontologies for representing objects in the context of this experiment and a more general manufacturing processing model which we detail in the following sections.

3.1 Ontology stack

To develop the ontology for the use case outlined above, we used a modified and shortened version of the method described by Uschold and King [26]. For the purposes of ontology development we used documentation containing domain-specific terminology and data as elicited from our colleagues at CAFHS. Fig. 1 shows the stack of ontologies we reused and developed within this project. The figure also includes example classes that are defined within each of these ontologies, whereas the dashes denote the subsumption relations between the classes. In the following sections we describe the classes and relations in these ontologies in more detail.

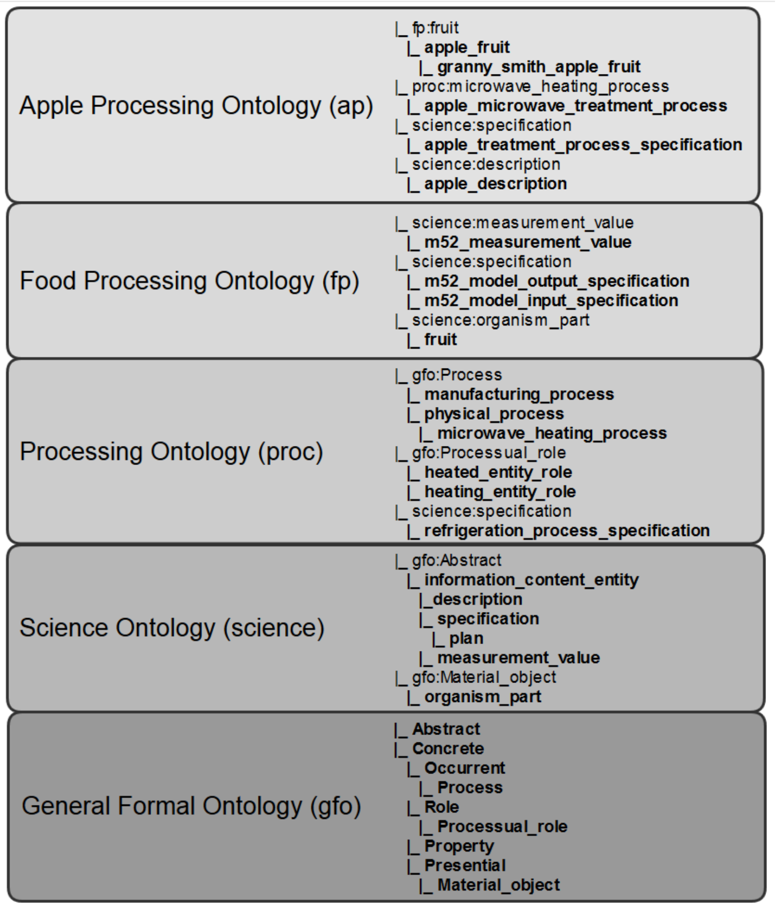


Fig. 1. Ontology Stack

Upper Level Ontology – GFO Upper– or "top level" or "foundational" –ontologies are ontologies of the most common entities in the world which are the same across all knowledge domains. For example, an upper ontology will provide a notion of what a material entity is, how material entities participate in processes and persist in time. The main purpose of an upper ontology is to facilitate semantic interoperability. A number of upper level ontologies are in use across the semantic Web community, though many Web ontologies are developed without referencing top level ontologies, often reducing the level of interoperability. Some of the most common ontologies currently in use across the semantic Web are DOLCE [18], the General Formal Ontology (GFO) [10], the Basic Formal Ontology (BFO) [8] and Cyc [14]. For the purposes of the work in this project, the General Formal Ontology was chosen, mainly for two reasons, (1) its well developed integration of objects and processes and (2) its well developed notions of time. Specifically, the GFO makes an explicit distinction between endurants (objects) and perdurants (processes) and provides convenient mechanisms for modelling how objects participate in processes. Time is taken to be primitive and time points (known as "time boundaries") can be derived. These time points can coincide which is useful for the modelling of continuous processes and change.

Science Ontology The Science Ontology [1] is a small ontology of terms which are common across all of physical science and engineering and resides underneath the General Formal Ontology. Typical terms contained in the ontology are "Information Content Entity", "Description", "Specification" including appropriate subterms. These are important for the disambiguation of the actual processes from process specifications, such as processing conditions etc..

Information Content Entities The most relevant concept that we reuse from the Science Ontology is that of an "Information Content Entity" (ICE) that is required to capture data, specifications and descriptions. ICEs are best described as entities that do not have independent existence, but rather are dependent on other entities and are in an "about-ness" relationship with those entities [5]. An ontological analysis of ICEs would conclude, that within the framework of the General Formal Ontology, these are subclasses of the $gfo{:}Abstract$ class. Abstract entities are entities which are independent from time and space, but may be dependent on other entities for their existence. Subclasses of information content entities that we reuse are, for example, "description", "measurement value" and "specification".

While perdurants such as processes may map onto a time vector, the entity that we observe when we measure time, for example, is not the time vector itself, but a representation, or in better terms, a descriptions of the time vector. In our ontology, processes therefore have "process specifications" which are in an aboutness relationship to the process itself – process specifications are types (subclasses) of descriptions, which, in turn, are information content entities. We may write:

1. $science{:}process\_specification\ subclassOf\ specification$
2. $science{:}process\_specification\ equivalentClassOf$ $(science{:}specification\ and\ (about\ only\ gfo{:}Process))$

Processing Ontology The microwave-based heat treatment process conducted in our use case is some kind of a food treatment process. Consequently, we need an ontology that defines concepts and relations of food treatment processes. However, to the best of our knowledge no such food treatment ontology exists. Defining such an ontology requires us to properly layer it on top of an Upper Level Ontology such as GFO. To do that we need to identify what constitutes a food treatment process and what are the more general concepts and relations that are needed to describe such a process. Looking at the concept of a food treatment process, it obviously involves some kind of "treatment process" that is performed on "food". A "treatment process" is performed either by a "human" or a "machine". The GFO makes a fundamental distinction between "Processes" and "Actions": ontologically, both are viewed as being of type "Occurrent" but are distinguished from each other through the involvement of an "Agent", i.e. an entity, playing an agent role. Agent roles can be played by both humans and machines. The entity that is treated during the treatment process is at its most general a *gfo:Material_object*. "Food" is – ontologically speaking – a role: a material object "becomes food" when it realises the food role (an apple sitting on a shelf, for example, is not "food" as it does not realise the food role).

From this very brief ontological analysis it becomes clear that we first need an ontology that describes objects in a factory, the roles these objects play as well as the processes in a factory and the mode of participation in those processes. A number of such manufacturing/processing ontologies exist already, however, they are either not layered on top of an Upper Level Ontology [15, 13] or do not provide the detail that we require from the processing ontology to model our use case [3]. We therefore developed a general purpose processing ontology, drawing inspiration from the referenced manufacturing ontologies as well as the United States patent and trademark offices taxonomy on manufacturing. As depicted in Fig. 1 the processing ontology, denoted by the namespace prefix "proc" in Fig. 1, defines, for example, different types of manufacturing processes, physical, chemical and biological processes.

Processes in the processing ontology are layered on the notion of processes in GFO which are characterised by the manner in which entities participate in them, i.e.:

1. *gfo:Process subclassOf gfo:Occurrent*
2. *gfo:Process subclassOf* (*gfo:has_role some gfo:Processual_role*)
3. *gfo:Processual_role subclassOf* (*gfo:role_of some gfo:Process*)

We defined a set of such roles that are common in manufacturing processes and that are played by material objects, such as a "heated_entity_role" and a "heating_entity_role". We may write:

1. *proc:heating_process subclassOf gfo:Process*
2. *proc:heating_process subclassOf* (*gfo:has_role some proc:heated_entity_role*)
3. *proc:heating_process subclassOf* (*gfo:has_role some proc:heating_entity_role*)
4. *proc:heated_entity_role subclassOf* (*gfo:role_of some gfo:Process*)
5. *proc:heating_entity_role subclassOf* (*gfo:role_of some gfo:Process*)

Further, we defined "chemical_material_objects", "physical_material_objects" and "biological_material_objects" that manufacturing processes take as input or produce as output such as a "machine", an "assembly_entity" and different types of "substances".

Processes can have other processes as part which allows the modeling of complex manufacturing processes and their breakdown into small process parts. Any other process can be modeled by analogy. Processes in our ontology have time boundaries with discrete start and end timepoints. The time boundaries for processes are mapped to the notion of “Chronoids” in GFO. Time is understood in GFO to be Brentano time [4]. The GFO defines “Chronoids” not as sets of points, but as entities in their own right, which have two outer and an infinite number of inner “time boundaries” [10]. Time boundaries can overlap which allows the modeling of continuous change. Processes project to a chronoid via the $\mathit{gfo{:}projects\_to}$ vector (relation):

1. $\mathit{gfo{:}Process\ subclassOf\ (gfo{:}projects\_to\ some\ gfo{:}Chronoid)}$
2. $\mathit{gfo{:}Chronoid\ subclassOf\ (gfo{:}has\_time\_boundary\ some\ gfo{:}Time\_Boundary)}$

This provides all the mechanisms needed to define process durations as well as start and end times and dates.

Food Processing Ontology On top of the manufacturing processing ontology we have developed a generic food processing ontology, denoted by the namespace prefix “fp” in Fig. 1 and a more use case-specific apple processing ontology, denoted by the namespace prefix “ap”. To the best of our knowledge, there exist no such ontologies, but for an improved interoperability we have included equivalence relations to the NCI Thesaurus[1] for all the biological concepts that are defined in the food processing ontology and apple processing ontology.

Material Objects Much of the use case experiment is concerned with apple processing and hence, apples can serve as an illustration of how we handle material objects in the ontology. As discussed above, the term “apple” is potentially polysemous and hence, we need to distinguish between an “apple tree” and an “apple fruit”. Furthermore, “apple fruit” must be subdivided into several types of apples such as “apple fruit on tree”, “harvested apple fruit” or “refrigerated apple fruit” if we wish to talk about fruit still ripening on trees as opposed to harvested ones and ones which have undergone some treatment. Ontologically speaking, all of these entities are subclasses of the GFO’s “material object” class. We may therefore write in First Order Logic:

1. $\mathit{ap{:}fruit\ subclassOf\ ap{:}organism\_part}$
2. $\mathit{ap{:}organism\_part\ subclassOf\ gfo{:}material\_object}$
3. $\mathit{ap{:}apple\_tree\ subclassOf\ ap{:}maleae}$
4. $\mathit{ap{:}malae\ subclassOf\ gfo{:}material\_object}$
5. $\mathit{ap{:}apple\_fruit\ subclassOf\ ap{:}fruit}$
6. $\mathit{ap{:}apple\_fruit\_on\_tree\ equivalentClassOf}$
   $\mathit{(ap{:}apple\_fruit\ and\ (part\_of\ some\ ap{:}apple\_tree))}$

Processes The ontological treatment of processes in the food processing ontology is analogous to the description outlined above: the apple microwave treatment process has at least three discernible participants and distinct roles: (a) an apple playing the role of the heated entity, (b) the microwave oven playing the role of the treating

[1] http://ncit.nci.nih.gov/

entity and (c) an apple playing the role of the treated entity. The apple in (a) and (c) are ontologically distinct: with the beginning of treatment process, the apple has ceased to be an untreated apple and become a treated apple. In first order description logic, we may define the treatment process as follows:

1. $\mathit{ap}{:}\mathit{apple\_treatment\_process}\ \mathit{equivalentClassOf}$
   $(\mathit{fp}{:}\mathit{microwave\_heating\_process}\ \mathit{and}\ (\mathit{proc}{:}\mathit{has\_participant}\ \mathit{some}$
   $(\mathit{ap}{:}\mathit{apple\_fruit}\ \mathit{and}\ (\mathit{gfo}{:}\mathit{plays\_role}\ \mathit{some}\ \mathit{proc}{:}\mathit{heated\_entity\_role})))\ \mathit{and}$
   $(\mathit{proc}{:}\mathit{has\_participant}\ \mathit{some}$
   $(\mathit{proc}{:}\mathit{microwave\_oven}\ \mathit{and}\ (\mathit{gfo}{:}\mathit{plays\_role}\ \mathit{some}\ \mathit{proc}{:}\mathit{heating\_entity\_role})))\ \mathit{and}$
   $(\mathit{has\_participant}\ \mathit{some}$
   $(\mathit{apple\_fruit}\ \mathit{and}\ (\mathit{gfo}{:}\mathit{plays\_role}\ \mathit{some}\ \mathit{proc}{:}\mathit{treated\_entity\_role}))))$
2. $\mathit{ap}{:}\mathit{apple\_microwave\_treatment\_process}\ \mathit{subClassOf}$
   $\mathit{fp}{:}\mathit{microwave\_heating\_process}$
3. $\mathit{fp}{:}\mathit{microwave\_heating\_process}\ \mathit{subClassOf}\ \mathit{gfo}{:}\mathit{Process}$

4 Information services

For our proof-of-concept service implementation to analyse and verify the experimental data we have chosen to use the SADI–semantic Automated Discovery and Integration–framework [27]. SADI is a semantic Web service framework that is predominantly used in the bioinformatics domain. SADI comprises a set of semantic Web compliant conventions and suggested best-practices for data representation and exchange between Web services. In contrast to other semantic Web service frameworks, SADI takes some assumptions that make the protocol and the implementation much easier than, for example, OWL-S [17] and WSMO [22]/WSMX [9]. SADI Web services are stateless, transformative, atomic and idempotent. The distinguishing simplification in SADI is that the input and output of a Web service must share a common "base" identifier, thus assuming that all services are "annotator services", where the Web services consume some specific input data type, and return a related output data type generated by whatever operation the service executes. Although our use case and the manufacturing domain typically require a process model to execute non-atomic manufacturing processes, we have chosen to use SADI for our first implementation for its ease-of-use. Further, SADI services are encapsulated functionalities that can be accessed over the HTTP protocol and thus, can, in the long run, be incorporated into a framework that allows for the execution of composite processes. Currently, we implement the process logic of composite processes in Java.

4.1 Architecture

We deployed our SADI information services onto an Apache Tomcat Server and use the Jena library to query the RDF Triple Store running on the same Tomcat instance (see Fig. 2). The RDF database is loaded with the ontologies as described in Sect. 3.1 constituting the *TBox*, while the actual data (the *ABox*), such as the temperature measurements for apples undergoing the heat treatment from different runs of the experiment, are loaded into the Triple store via scripts that transform the raw sensor data into ontology instances. The services use SPARQL queries to retrieve the required information from the Triple store, process them and return the annotated ontology instance back to the client.

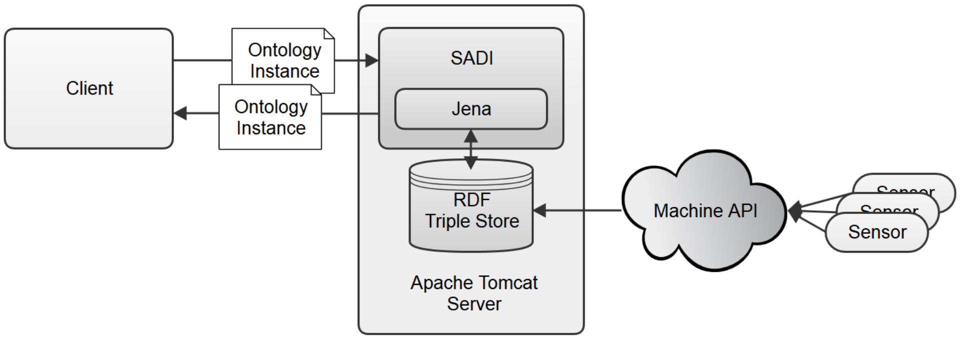


Fig. 2. System architecture of our semantic lab Notbook

4.2 Experimental Constraints Compliance service

The first semantic Web service we developed for analysing data in the semantic lab notebook allows to verify if an apple undergoing the treatment process was refrigerated properly before the experiment. The service takes as input (see Listing 1.1) a specific refrigeration process specification and gives a boolean return (see Listing 1.2) confirming if all the temperature observations recorded for the apples participating in the refrigeration process comply to the limits defined in the specification. The refrigeration process itself references the individual apples/batches through a *gfo:plays_role* relation. The semantic Web service uses SPARQL to query the specification of the provided process and to extract the allowed minimum and maximum temperature values for each refrigeration process. Then all temperature observations for apples playing a role in the given process are extracted and checked against the specified limits. A boolean attribute *fp:isCertifiedProcess* is then added to the instance of the process specification which is in turn returned by the service.

A false return value would indicate an interrupted cooling chain and thus nullify the results of the experiment.

Listing 1.1. Input RDF

```
<!DOCTYPE rdf:RDF
   [ <!ENTITY matinf ”http://matinf.cmse.csiro.au/”> ]>
<rdf:RDF xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
   xml:base=”@matinf;id/exp_12-09-12#”
   xmlns:fp=”@matinf;ont/owl/processing.owl#””>
  <owl:Ontology rdf:about=”@matinf;id/exp_12-09-12#”/>
  <fp:refrigeration_process_specification
     rdf:about=”@matinf;id/exp_12-09-12#refrigeration_proc_spec_1”>
  </fp:refrigeration_process_spec>
</rdf:RDF>
```

Listing 1.2. Output RDF

```
<!DOCTYPE rdf:RDF
   [ <!ENTITY matinf ”http://matinf.cmse.csiro.au/”> ]>
<rdf:RDF xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
```

```
  xml:base=”@matinf;id/exp_12-09-12#”
  xmlns:fp=”@matinf;ont/owl/foodprocessing.owl#”>
  <fp:refrigeration_process_specification
      rdf:about=”@matinf;id/exp_12-09-12#refrigeration_proc_spec_1”>
    <fp:isCertifiedProcess
        rdf:datatype=”http://www.w3.org/2001/XMLSchema#boolean”>
      true
    </fp:isCertifiedProcess>
  </fp:refrigeration_process_specification>
</rdf:RDF>
```

This service essentially implements a complex SPARQL query using FILTERS and thus could also be expressed in a SPARQL templating language such as SPIN [12] and then executed on demand. However, the next service uses complex calculations that cannot be expressed in SPARQL. Thus, it represents one of many computations in our use case that require algebraic computations that cannot be expressed in SPARQL or RDFS/OWL directly.

4.3 Continual M_{52} computational service

The second service expects an instance of an “apple_description” as input (see Listing 1.3) and returns a M_{52} time equivalent [25] (e.g. “m52_model_output_specification_1” in Listing 1.4) that has been achieved during a microwave treatment process for the given apple (described by the apple description). The returned “m52_model_output_specification_1” instance is a specification itself while the actual value that was created for the M_{52} time equivalent can be queried via the following SPARQL query:

```
SELECT ?o
WHERE { <base:m52_model_output_specification_1> <science:has_Value_Literal> ?o }
```

In our case this query returns a value of 22.62 minutes which indicates how long the embedded larvae in the specific apple would have been exposed to a reference temperature of $52^\circ C$ in the treatment process. To calculate this value, the service retrieves all temperature observations for each of the four sensors embedded in the specific apple described by “granny_smith_apple_desc_001” and incrementally accumulates the minimum accumulated total temperature equivalent to M_{52}. This value indicates if the apple was sufficiently heat treated in the experiment to kill all embedded larvae.

Listing 1.3. Input RDF

```
<!DOCTYPE rdf:RDF
  [ <!ENTITY matinf ”http://matinf.cmse.csiro.au/”> ]>
<rdf:RDF xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
  xml:base=”@matinf;id/exp_12-09-12#”
  xmlns:ap=”@matinf;ont/owl/appleprocessing.owl#”>
  <owl:Ontology rdf:about=”@matinf;id/exp_2012-09-12#”/>
  <ap:apple_description
    rdf:about=”@matinf;id/exp_12-09-12#granny_smith_apple_desc_001”>
  </ap:apple_description>
</rdf:RDF>
```

Listing 1.4. Output RDF

```
<!DOCTYPE rdf:RDF
  [ <!ENTITY matinf "http://matinf.cmse.csiro.au/"> ]>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xml:base="@matinf;id/exp_12-09-12#"
  xmlns:ap="@matinf;ont/owl/appleprocessing.owl#"
  xmlns:science="http://purl.org/scimantica/owl/science.owl#">
  <ap:apple_description
    rdf:about="@matinf;id/exp_12-09-12#granny_smith_apple_desc_001">
      <science:is_specified_by
        rdf:about="@matinf;id/exp_12-09-12#m52_model_output_spec_1"/>
      </science:is_specified_by>
  </ap:apple_description>
</rdf:RDF>
```

5 Related Work

This is not the first time that semantically enabled electronic lab notebooks have been proposed [11, 21, 7]. Many of the related works describe the techniques and the methodologies on how to introduce metadata to improve the provenance of experiments. For example, in [24] it is proposed that a widely used lab notebook, ELN, be extended with semantic annotation capability to support integration with external annotation sources such as produced by problem solving environments. Other works have gone farther and actually published the data of experiments in RDF/OWL [6].

Some others have developed tools focussed on different aspects of the experimental data curation problem, for example on the scalability [23], or on capturing the relationships between results of different experiments [19].

[20] proposes an aggregation tool based on RSS feeds to ensure that the objects created during the research process are recognized, stored and indexed.

The bioinformatics community as a whole is spearheading other academic disciplines by capturing vast quantities of the knowledge published in scientific articles as ontologies in the Bioportal initiative[2].

However, we are not aware of any prior works on capturing the data produced by an experiment in ontologies combined with custom-built RESTful semantic Web services on top of the RDF data that allow to reproduce and verify the results of the experiment. The closest work to ours, but with a stronger focus on capturing the entire workflow of an experiment was proposed in [16]. The work introduces a laboratory domain specific ontology and the COW (Combining Ontologies with Workflows) software tool was developed to formalize workflows which were enhanced with ontological concepts taken from the developed domain specific ontology.

6 Conclusion

We described our prototypical implementation of a semantic lab notebook that allows data obtained in an experiment to be stored in RDF and accessed via SPARQL and custom-built semantic Web services. The system allows scientists to read the experiment related data and to combine it as part of a scientific workflow. We implemented ontologies required to model our data points via Protégé in OWL and

[2] http://bioportal.bioontology.org/

developed SADI RESTful Web services in Java that implement computational analysis functionality that cannot be expressed in the ontology directly. With an increased availability of ontologies and tools that support the capture of RDF, semantic lab notebooks can play a significant role in helping the research community to store experiment data consistently, process it faster and allow the mashup of collected datasets to facilitate research studies over multiple datasets, while reducing complexity and error rates.

Acknowledgements We are thankful to the support of our colleagues in the CSIRO Animal, Food and Health Sciences division, in particular to Dr. Mala Gamage who regularly shared the results of their heat-treatment experiments with us.

References

1. N. Adams. Science Ontology. `https://github.com/scimantica/science-ontology`, 2013. [Online; accessed 12-July-2013].
2. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 2001.
3. S. Borgo and P. Leitão. Foundations for a Core Ontology of Manufacturing. In R. Sharman, R. Kishore, and R. Ramesh, editors, *Ontologies*, volume 14 of *Integrated Series in Information Systems*, pages 751–775. Springer US, 2007.
4. F. Brentano. *Philosophische Untersuchungen zu Raum, Zeit und Kontinuum.* Felix Meiner Verlag, Hamburg, 1976.
5. W. Ceusters. An information artifact ontology perspective on data collections and associated representational artifacts. *Studies in health technology and informatics*, 180:68—72, Jan 2012.
6. J. Frey, D. Roure, K. Taylor, J. Essex, H. Mills, and E. Zaluska. CombeChem: A Case Study in Provenance and Annotation Using the Semantic Web. In L. Moreau and I. Foster, editors, *Provenance and Annotation of Data*, volume 4145 of *Lecture Notes in Computer Science*, pages 270–277. Springer Berlin Heidelberg, 2006.
7. J. G. Frey. The value of the Semantic Web in the laboratory. *Drug Discovery Today*, 14(1112):552 – 561, 2009.
8. P. Grenon and B. Smith. SNAP and SPAN: Towards dynamic spatial ontology. *Spatial cognition and computation*, 4(1):69–104, 2004.
9. A. Haller, E. Cimpian, A. Mocan, E. Oren, and C. Bussler. WSMX – A Semantic Service-Oriented Architecture. In *International Conference on Web Services*, pages 321 – 328, Orlando, FL, USA, july 2005. IEEE Computer Society.
10. H. Herre, B. Heller, P. Burek, R. Hoehndorf, F. Loebe, and H. Michalek. General formal ontology (gfo)–a foundational ontology integrating objects and processes. *Onto-Med Report*, 8, 2006.
11. G. Hughes, H. Mills, D. De Roure, J. Frey, L. Moreau, M. Schraefel, G. Smith, and E. Zaluska. The semantic smart laboratory: A system for supporting the chemical eScientist. *Organic and Biomolecular Chemistry*, 2:3284–3293, 2004.
12. H. Knublauch, J. A. Hendler, and K. Idehen. SPIN – SPARQL Inferencing Notation. http://www.w3.org/Submission/2011/SUBM-spin-overview-20110222/, 2009.
13. S. Lemaignan, A. Siadat, J. Y. Dantan, and A. Semenenko. MASON: A Proposal For An Ontology Of Manufacturing Domain. In *IEEE Workshop on Distributed Intelligent Systems: Collective Intelligence and Its Applications*, pages 195–200, 2006.
14. D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

15. H. Lin and J. Harding. A manufacturing system engineering ontology model on the semantic web for inter-enterprise collaboration. *Computers in Industry*, 58(5):428–437, 2007.
16. A. Maccagnan, M. Riva, E. Feltrin, B. Simionati, T. Vardanega, G. Valle, and N. Cannata. Combining ontologies and workflows to design formal protocols for biological laboratories. *Automated Experimentation*, 2, 2010.
17. D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara. OWL-S: Semantic Markup for Web Services. Member submission, W3C, 2004.
18. C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, R. Oltramari, L. Schneider, L. P. Istc-cnr, and I. Horrocks. Wonderweb deliverable d17. The Wonderweb library of foundational ontologies and the DOLCE ontology. 2002.
19. J. Myers, C. Pancerella, C. Lansing, K. Schuchardt, and B. Didier. Multi-Scale Science, Supporting Emerging Practice with Semantically Derived Provenance. In *ISWC workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, 2003.
20. C. Neylon. Head in the clouds: Re-imagining the experimental laboratory record for the web-based networked world. *Automated Experimentation*, 1, 2009.
21. E. Polonsky, E. Polonsky, A. Six, M. Kotelnikov, V. Polonsky, R. Polly, and P. Brey. Semantic laboratory notebook: Managing biomedical research notes and experimental data. In *Proceedings of ESWC*, 2006.
22. D. Roman, U. Keller, H. Lausen, J. de Bruijn, R. Lara, M. Stollberg, A. Polleres, C. Feier, C. Bussler, and D. Fensel. Web Service Modeling Ontology. *Applied Ontologies*, 1(1):77–106, 2005.
23. M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan, and S. Xu. Data curation at scale: The data tamer system. In *CIDR*, 2013.
24. T. Talbott, M. Peterson, J. Schwidder, and J. D. Myers. Adapting the electronic laboratory notebook for the semantic era. In *Proceedings of the 2005 international conference on Collaborative technologies and systems*, CTS'05, pages 136–143, Washington, DC, USA, 2005. IEEE Computer Society.
25. J. Tang, J. Ikediala, S. Wang, J. Hansen, and R. Cavalieri. High-temperature-short-time thermal quarantine methods. *Postharvest Biology and Technology*, 21(1):129 – 145, 2000.
26. M. Uschold and M. King. Towards a Methodology for Building Ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, 1995.
27. M. Wilkinson, B. Vandervalk, and M. L. The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. *Journal Biomed Semantics*, 2, 2011.

BiographyNet: Managing Provenance at multiple levels and from different perspectives

Niels Ockeloen, Antske Fokkens, Serge ter Braake, Piek Vossen, Victor de Boer, Guus Schreiber, and Susan Legêne

The Network Institute, VU University Amsterdam
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{niels.ockeloen,antske.fokkens,s.ter.braake,piek.vossen
v.de.boer,guus.schreiber,s.legene}@vu.nl
http://wm.cs.vu.nl

Abstract. The BiographyNet project aims at inspiring historians when setting up new research projects. The goal is to create a semantic knowledge base by extracting links between people, historic events, places and time periods from a variety of Dutch biographical dictionaries. A demonstrator will be developed providing visualization and browsing techniques for the knowledge base. In order to establish its credibility as a serious research tool, keeping track of provenance information is crucial. This paper describes a schema that models provenance from different perspectives and at multiple levels within BiographyNet. We will present a concrete model for the BiographyNet demonstrator that uses elements from the Europeana Data Model [6], PROV-DM [17] and P-PLAN [11].

Keywords: eHumanities, Linked Data, PROV-DM, P-PLAN, ORE, EDM

1 Introduction

E-humanities investigates what can be done in humanities with modern techniques which we could not do before, or only could do with a great deal of effort. E-history is a subdomain of e-humanities which offers a way of linking pieces of information and discovering relationships which otherwise would be difficult to trace. It generally aims at improving methods of existing historical research rather than introducing a whole new way of historical research [22]. It creates pathways through information, rather than being the closing factor or end result in historical research [1, 41]. Efforts in e-humanities often concentrate on how to mine 'big data', which we define as data which is very difficult to handle manually for a traditional researcher. More challenging, and in general also more interesting, are projects which aim to go beyond the simple data mining and endeavor to answer difficult research questions like the similarity between and interdependability of two or three texts, tracing and defining the subjective elements and descriptions, or signaling traces of political or cultural influences from a society during a given period. These new ways of mining historical data lead to new questions on provenance of information. It is imperative for historians to keep a good oversight over the sources which were used to produce a certain output. How reliable are the sources which were used and what do they tell about the significance of the outcome? What differences are found in the information that individual sources provide? When information differs, how are specific points of view distributed over different sources? How can results be manipulated by adjusting queries for a more accurate result? For these reasons, the historian needs to have an aggregated view of the process from query to output and, if necessary, inspect the whole process step by step to learn which additional sources and heuristics were involved.

1.1 Use Case: BiographyNet

The BiographyNet project is an e-history project bringing together researchers from history, computational linguistics and computer science. The project uses data from the Biography Portal of the Netherlands (BP), which contains approximately 125,000 biographies from a variety of Dutch biographical dictionaries, describing around 76,000 individuals. The aim of BiographyNet is to develop a demonstrator which supports the discovery of interrelations between people, events, places and time periods in biographical descriptions. Through a combination of data enrichment, quantitative analysis, visualization and browsing techniques, the demonstrator should provide leads and insights that may be hard to discover using traditional methods. As such, it may inspire historians to investigate more ambitious research questions.

The BP links biographies written by thousands of authors with very different temporal and academic backgrounds. This results in many levels of reliability of the 125,000 entries in this melting pot of Dutch biographies. Provenance information is therefore an important factor. It must however be noted that provenance information on the original sources does not go beyond the information that is provided by the BP such as author, publisher or the book from which a text was taken.

2 Motivation

The demonstrator should help historians do their research. This goal can only be met if the validity of the demonstrator's results can be verified. To this end, information needs to be available on performed operations as well as on used sources. According to Groth et al. [12], "data can only be meaningfully reused if the collection processes are exposed to users. This enables the assessment of the context in which the data was created, its quality and validity, and the appropriate conditions for use". Hence, provenance plays an important role in establishing the demonstrator's credibility.

Provenance needs to be modelled from different perspectives and at multiple levels for BiographyNet. These different perspectives include 1) the perspective of the information used to produce the results provided by the demonstrator, e.g. which original sources contributed to the outcome, 2) the perspective of the processes involved in creating the results and 3) the perspective of the people that were involved in setting up the pipeline of processes. The various levels include 1) provenance at component level, recording each aspect of the processing steps involved such as tool name, version, etc. and 2) an aggregated view of the provenance information for the interlinked processes as a whole. The latter is targeted at the end user of the system, in this case the historian, while the former is needed by the computer scientist in case the outcome of an aggregated process is pulled into question.

In the next two sections, we address the requirements for provenance modelling specific for BiographyNet. First, we will address the point of view of historians who are primarily interested in the reliability of the system. We will explain how the requirements for historians relate to the categories for provenance on the web defined by Groth et al [12]. Section 4 will outline BiographyNet from the point of view of the system developers whose primary interest is to improve the technology behind the demonstrator. Section 6 will describe the BiographyNet schema devised to allocate the required provenance information as described in the preceding sections.

3 Requirements for Historians

There are two main requirements for the historian regarding provenance when using the demonstrator: A trace back to the text and metadata in the original source, and insight into the processes manipulating and selecting the original data. We will explain the first requirement through a research question on the background of the 71 governors-general of the Dutch Indies between 1610 and 1949. If, for instance, we run a query to find out what the average age of these individuals was at the time of their appointment, provenance information of different granularity should be present: a) an overview of the sources (in our case biographical dictionaries) that were used for the overall outcome and how often each individual source was consulted, b) an overview of potentially relevant data that was excluded from the end result. This is important in case of conflicting data, where one source generally considered more reliable was used rather than another and c) the sources that were used for a specific results (i.e. the age of a specific governor at the time of his appointment).

One can assume that few historians will have the background to completely (or even partly) understand the finer technical details of how data are processed in order to answer a query. Even when a new generation of ‘e-historians’ is trained, one cannot expect them to be computer scientists. Therefore, provenance of data manipulation should be modelled as simple as possible and focus on aspects that may directly influence the outcome of research questions. First and foremost, it should always be indicated whether information is directly extracted from the metadata or the result of automatic interpretation of text. Complete accuracy in automatic text interpretation cannot be guaranteed. Information extracted from text should therefore always include a direct link to the original source. Provenance should also indicate the overall performance of the system that interpreted the text; depending on the kind of question, the historian may want to have results that aim for high recall or high precision. Finally, a global description of heuristics used when interpreting data should be provided. While resolving ambiguous location names, for instance, a strategy that always prefers locations in or near the Netherlands is likely to lead to good results within the BiographyNet project. However, if the historian wants to investigate the ties between officials in the former Dutch colonies (where cities with Dutch names can be found), this strategy would bear a direct undesirable influence on the results. The historian should thus be able to check whether the interpretation process used any strategies that may introduce a bias that influences results.

If we translate this to the categories outlined in [12], this leads to the following requirements.[1] The **objects** for which we need to model provenance are texts from several sources, metadata and statements extracted from the text. Texts and metadata are **attributed** to publishers and authors of this data. Extracted information should also indicate the author or publisher of the original text and, in addition, point to the system used to extract the information. There is thus a tight link between the process and the attribution while modelling provenance of automatically extracted text. Attribution plays a significant role in establishing the reliability of information and this includes the reliability of the methods that were used to extract information from text.

Information on the **process** should include detailed indications of the system?s overall performance: i.e. it should indicate the precision and recall of the system for specific categories. Furthermore, the **version**, publication date and person **responsible** for generating the output should be indicated in case the historian wants to replicate their results at a later stage. Finally, provenance should include **justifications** for decisions made in the extraction process, in particular concerning techniques used to disambiguate terms or resolve entities. The historian may need such information to check whether the information extraction used heuristics or forms of **entailment** that may interfere with the outcome of the research question addressed by the demonstrator, as illustrated by the location disambiguation example above.

In order to address the aspects of **trust** and **accountability** as outlined above, it must be crystal clear which information comes directly from original sources, and which information is the result of the processing or interpretation of these sources. Hence, the schema for BiographyNet should accommodate for this. The distinction should be marked prominently, because automatic processes add a dimension to reliability that not all historians will be familiar with. One of the main challenges therefore is that technical processes should be explained in terms that are understable to researchers who generally do not have a technical background. Strong collaboration between the historians and system designers is thus required when designing this part of provenance modelling throughout the project. At this level, an indication of responsibility is necessary so that

[1] Concepts that are addressed in [12] will be marked in **bold font**.

historians can contact the persons who designed the interpretation pipeline in case of an unexpected outcome or if questions arise on the made assumptions or used heuristics.

4 Requirements for computer scientists

Researchers working on demonstrators are mainly interested in provenance because it helps to make experiments replicable and it supports research to improve existing technologies. We use the term **replication** to refer to the process of following the exact same procedure as in the original work and thereby obtain the exact same output. This is different from reproduction where the same question is answered using different means (e.g. a new implementation or evaluation set). The validity of research results increases when they can be reproduced, whereas replication only verifies that an outcome was valid under specific conditions [8]. Within our setup, replication matters for two reasons. First, we need to be able to create the exact same dataset for historians if they want to compare new results to previous results. Second, when results cannot be reproduced, it is almost impossible to find the cause without being able to replicate the original results [18].

It is well known that both replicating and reproducing results is challenging when computer programs are involved. This especially holds if the code is not available [19, 18] but even if code is present [21, 10]. Fokkens et al. [10] define five categories that may influence results in pipelines that involve Natural Language Processing (NLP). They are preprocessing (e.g. tokenization, cleaning up data), experimental setup (e.g. splitting folds for 10-fold cross validation, evaluation set), versioning (e.g. version of resources such as WordNet [9], or tools such as Mallet [15] for machine learning), system output (e.g. the exact features for specific tokens, intermediate output of the system in a pipeline) and system variation (e.g. treatment of ties, thresholds). This information must be explicit in order to replicate results.

Information on influential factors immediately contributes to the second use of provenance for computer scientists: improving existing technologies. Individual tools and datasets interact in different ways with each other. Systematic testing of influential parameters, exchanging tools for subtasks and combining the output of different tools can lead to significant improvement in performance. The interaction between performance of subtasks and overall performance of the system is not always straightforward. The output of the sentence splitter, for instance, influences the output of the parser. However, even if the output of the parser of the utterance as a whole is incorrect, we may still obtain the grammatical relations we need to identify the participant of an event.

The **object** for which we need to model provenance thus is the data at various stages of the provenance pipeline. This data is **attributed** to a specific tool that has taken data from the previous stage and possibly one or more external resources as input. Again, attribution is tightly linked to the **process**. Modeling the process is the most complex aspect of modelling provenance for the NLP pipeline. It requires registering detailed information on all tools and data sets involved including preprocessing steps, steps to generate features and the process of creating training data for machine learning. For all tools and resources, the **version** should be indicated. A detail in implementation or a small step or setting can make a significant difference in the results. It should therefore be registered who is **responsible** so that differences can be traced when third parties do not manage to reproduce results. Finally, documentation should clearly describe the decisions made in the setup which both serves as a **justification** of the approach and a way to indicate any form of **entailment** that may be required by the historian.

5 Retrieving information from text

One of the main challenges of building a demonstrator lies in creating tools that can automatically interpret text and extract information from it. The design of the system that is responsible for automatic text interpretation is work in progress. We will therefore provide a description of what this process is likely to look like based on the work carried out so far as well as systems used in related work. The main purpose of this section is to provide an indication of the different steps involved in automatic text interpretation.

We start by identifying linguistic information in text, where we distinguish two processes: named entity recognition and concept identification. Named entity recognizers identify names of persons, organizations and locations. Some also identify dates. We will use an off-the-shelf named entity recognizer for Dutch, for instance LingPipe[2]. Concept identification involves linking words in text to a set of concepts of interest. We will use revisions of tools described in [20] and [5]. Their approach is based on McCarthy et al's [16] observation that words tend to have a predominant sense within a specific genre or domain. The approach involves two steps. Concepts of interest are first identified in the corpus whereafter an executing step is performed in which these concepts are labeled in the text. We will briefly describe the two steps below.

- First, candidate terms are identified in the text. In a basic system, these may be verbs and nouns co-occurring in a sentence. We thus start by running a sentence identifier, tokenizer and part-of-speech tagger and lemmatizer over the entire corpus.
- Next, we link all these terms to WordNet entries and create hypernym chains. This process results in an overview of the hypernym chains identified in the text. For each hypernym, the set of hyponyms occurring in the text is given. We manually select a set of hypernyms from this overview. This set of hypernyms constitutes our concepts of interest. As soon as we have created a set of concepts of interest, we can tag these concepts in the text. First, we create a corpus by running a tokenizer, part-of-speech tagger and lemmatizer over the text. For each lemma in the corpus, we check whether

[2] `http://alias-i.com/lingpipe/web/demo-ne.html`

one of its senses is a hyponym of one of the concepts of interest. In this case, we associate the lemma to this concept of interest. Lemmas are thus only linked with selected concepts of interest and the senses that are related to these concepts constitute their predominant sense within our domain.

Together, named entity recognition and concept identification provide a corpus in which persons, organizations, times, locations and concepts are labelled.

Consequently, we can apply two strategies to extract useful information from text: a rule based approach and an machine learning approach (ML). We can define basic mapping rules that directly map the resulting labels within this corpus to usable metadata. If for instance, we encounter a person name identified by the named entity recognizer in close proximity of a profession tagged by our concept identifier, we assign this profession to the person.

The ML strategy uses existing metadata to discover similar information in biographies for which that metadata is missing using named entities and concepts as features. The biographies obtained from the BP are accompanied by metadata that includes information on the subject of the biography. The completeness of this metadata varies significantly from source to source. Biographies with rich metadata can be used to learn to identify information in text and hence find this information in biographies with poorer metadata. We have created a corpus in which information from metadata is tagged in the original text of the biography. This corpus can be used as a training set for machine learning to discover information in texts that is missing in the metadata. For example, we found that the metadata field 'religion' was available for only 6 out of the 71 governor-generals in our use case. However, using ML we found this information in the text for 20 governors.

Together these strategies form the core of our system for text interpretation. It should be noted that the descriptions provided above illustrate a basic system that is currently under development. Throughout the project, we will incrementally improve the system by adding more linguistic information.

6 The BiographyNet schema

Having outlined the main concerns and requirements for the BiographyNet demonstrator, the following section describes the schema devised to manage the data used and produced for the demonstrator. It describes how data from both original sources and enrichments is stored, how provenance information is handled for involved processes and how this ties into the formulated requirements. An impression of the schema can be found at: `http://www.biographynet.nl/schema/`. The following subsections are best read with the schema alongside. The mentioned concepts and relations can then be traced and followed in the schema. Description of the various parts of the schema generally takes place from left to right. Please note that this impression includes the various aspects described in this section in order to provide a general overview of the schema for BiographyNet. It does not include every aspect of the biographical data and provenance information in order to maintain overview. Information on individual Activities, Entities etc. such as start times, version numbers etc. is left out and qualified relations are only modelled if needed to illustrate the ideas behind the schema.

6.1 Foundations of the BiographyNet schema

The collection of biographies is made available to the BiographyNet project as a collection of XML files. Each XML file contains a 'Biographical Description', which in turn contains three different types of data; A 'File Description' that contains the metadata on the original source, a 'Person Description' that contains limited metadata on the depicted person, and the actual biographical description. Currently, the available biographical data is not linked to any other sources. To be more flexible when it comes to linking to external sources in the near future and in order to reason over the data, the BiographyNet demonstrator will be based on Linked Data [2] principles. Therefore, the collection of XML files is converted to RDF [4]. How this conversion was done in detail is out of scope for this paper, but a similar conversion process is described in [3]. When data needs to be converted, it is advisable to stay as close as reasonably possible to the original schema, in this case defined by the structure of the XML files. Any altering of the schema involves interpretation, and as interpretation can change over time, such a process has the potential for information loss. For this reason we started out with a schema for BiographyNet that closely follows the structure of the original XML files; it contains a resource that represents a 'Biographical Description' (BioDes) that has connections with resources that represent a 'File Description' (FileDes), a 'Person Description' (PersonDes) and a resource for 'Biographical Parts' (BioParts). In the illustration, these are the blue outlined ovals, starting with the second leftmost.

Within the provided collection, multiple biographical descriptions are often available for the same person, originating from different sources. While these are represented as separate XML files in the provided collection, they need to coexist within the created Linked Data corpus. To this end, the BioDes objects are tied together using a resource representing the depicted person. This is the leftmost blue outlined oval. However, this means that -through the BioDes objects- a person can have multiple PersonDes objects containing possibly conflicting sets of metadata. In order to make the semantics of this more clear, we used parts of the Open Archives Initiative's 'Object Re-use & Exchange' ontology (OAI-ORE) [13, 14] in a way similar to how the Europeana Data Model (EDM) [7] uses concepts from that ontology. By defining the PersonDes objects as a subclass of the ore:Proxy class, defining the depicted person as an edm:ProvidedCHO (Cultural Heritage Object) and incorporating the associated predicate relations, the model becomes compatible with the Europeana data model while still staying true to the original data structure. The depicted person can now be viewed as a 'Cultural Heritage Object', of which multiple sets of metadata are made available through proxies, indicating that these sets of metadata represent different 'views' of that person.

This solution also allows for adding a new BioDes object for a person that 'aggregates' multiple other sources (BioDes objects) through the ore:Aggregates and edm:AggregatedCHO predicates. Besides the original biographical descriptions and an aggregated version of them, the model can also be used to accommodate enrichments. In that sense, an enrichment is a ‘new’ biographical description which was *derived from* original sources. A FileDes object will not be available for the enrichment, as the enrichment itself does not *directly* come from an original source, i.e. a biographical dictionary. Similarly, a BioDes object for an enrichment will most likely not contain a BioParts object, as it represents a set of metadata resulting from the enrichment process, but does not contain actual biographical texts. By modelling the *was derived from* relation, the enrichment can be traced back to the biographical description it was derived from and its original source, a hard requirement formulated in section 3.

6.2 Extending the schema with Provenance

PROV-DM [17] is the logical candidate for modelling provenance, since W3C[3] made it a recommendation promoting its widespread use. Furthermore, PROV concepts can be modelled in RDF making it suitable for use in the BiographyNet schema. Besides relations such as *was derived from*, the PROV ontology can be used to model Entities, Agents and Activities that played a role during the enrichment process and the creation of the pipeline of processes itself, including their mutual relations. Additionally, concepts from the new P-PLAN [11] are integrated in the BiographyNet schema to specify plans made for the actual activities involved in the enrichment process. Specifying planning information is useful in that it provides a way of verifying to what extend actions were performed according to plan. Hence, integrating this information makes it easier to identify errors in individual processes of the aggregated enrichment process. It also makes replication of results more feasible, as the plans provide a description of what the input and output of activities should look like. As such, the combined use of these ontologies ties into the requirements of the historian to be able to trace which original sources were used to obtain a result and to gather additional information on possible heuristics and biases. It also ties into the requirement of the computer scientist to be able to replicate results. In order to fulfill the requirements of the historian and computer scientist to have both an aggregated view on provenance (i.e. which original sources contributed to an enrichment) and a detailed view (i.e. specified information for all processing steps involved), these two levels are modelled separately in the schema. In the illustration, the aggregated level is represented by the orange outlined ovals (and the green one for the plan) between the two blue biographical structures. The detailed view is made up by the remainder of the schema. Clearly visible in the schema is how these activities and plans are parts and steps of the aggregated enrichment activity and its associated plan. These two views are described in more detail in the subsections below.

[3] `http://www.w3.org/`

6.3 Aggregated provenance information

A prov:wasDerivedFrom relation is made between the BioDes object of the enrichment and the BioDes object of the original source in order to model the information on an enrichment process as a whole. Furthermore, a prov:Activity to represent the aggregated process and its relations to the BioDes objects are specified. That activity has a prov:Agent associated with it. This Agent is the aggregated set of tools used for the enrichment, otherwise known as a 'pipeline'. The desired behavior of the integrated process is described by a prov:Plan object, which has its own provenance information; the plan for the enrichment process is attributed to an Agent, e.g. a computer scientist and can be derived from an earlier version of that plan or another enrichment. This aggregated provenance view allows the end user to identify which enrichments were used to produce a final aggregated view of information. The end user can determine the original sources through the various provenance relations. Furthermore, the end user knows who to contact in case an enrichment process seems to have produced questionable results. The aggregated plan can provide an overview of the input variables used in the underlying processes, as they are referenced through p-plan:isVariableOfPlan relations. This information allows for possible adjustments in order to adjust the output of the overall process.

6.4 Detailed provenance information

The detailed provenance information on individual processes is modeled as a chain of Activities which all have their own input and output Entities, associated Agents and Plan. The Agents are specific tools such as a tokenizer or part-of-speech tagger. The plan describes what the specific tool should do. Each Plan has its own provenance information. These plans are plans in their own right, but are also designated a p-plan:Step to indicate that they are a step of the aggregated plan for the enrichment as a whole. As such, these steps have input and output variables that describe the input and output of the related Activity. These variables correspond to the entities used by and generated by the related activities. An Activity together with its used and generated Entities can be seen as a 'bundle' of objects that together are derived from the Plan for that activity. Each individual Activity is designated as a part of the aggregated enrichment Activity using the Dublin Core 'hasPart' predicate. The order in which the individual Activities are executed can be derived from the prov:used and prov:wasGeneratedBy relations that tie the individual Activities to the Entities representing intermediate results. Besides these intermediate results, other Entities may be used by a specific Activity, e.g. a list of cities for Named Entity Recognition. For both the intermediate results as well as these 'external sources', the data format is unknown. An intermediate result could be a collection of RDF triples, an XML file or plain text file. An external source could be one of those or basically any type of document. In order to cope with this variety, these Entities are represented by a prov:Entity of subclass bgn:IntermidiateResult or bgn:ExternalSource, that can point to the actual document or serve as Named Graph to contain RDF data.

The aggregated view and detailed view of provenance information are related together by the fact that all Activities in the detailed view are parts of the aggregated Activity, all Plans of the individual Activities are Steps in the aggregated Plan and the biographical description of the 'Source' BioDes object is the actual Entity that is used by the first individual Activity, whereas the Entity produced by the last individual Activity is the resulting set of metadata of the enrichment BioDes object. Any form of pre- or post-processing of input data or results, needed to relate to those objects, needs to be viewed as a separate individual step in the overall plan. For without provenance information on those steps, replicability is not ensured.

7 Conclusion

Keeping track of provenance information is essential for the BiographyNet demonstrator to be viewed as a valid research tool for historians. In this paper we described why this is the case, what the requirements are to model provenance from multiple perspectives and which existing ontologies we used to devise a schema for BiographyNet that meets those requirements. We presented a first version of the BiographyNet schema that not only models provenance on what has taken place, but also models plans to compare against. The next step is to proceed with building a first version of the demonstrator. We will then have to evaluate how the schema holds up in practice, and use the output of such evaluation to further improve the schema.

Acknowledgements

This work was supported by the BiographyNet project (Nr. 660.011.308), funded by the Netherlands eScience Center (http://esciencecenter.nl/). Partners in this project are the Netherlands eScience Center, the Huygens/ING Institute of the Royal Dutch Academy of Sciences and VU University Amsterdam.

References

1. Arthur, P.: Exhibiting history. the digital future. Recollections 1(1) (2008)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: ESWC, volume 7295 of Lecture Notes in Computer Science. p. 733?747. Springer Berlin / Heidelberg (2012)
4. Carroll, J.J., Klyne, G.: Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, W3C (Feb 2004), http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

5. Cybulska, A., Vossen, P.: Using semantic relations to solve event coreference in text. In: Mititelu, V., Popescu, O., (Eds.), V.P. (eds.) Proceedings of the Workshop Semantic relations-II. pp. 60–67. Istanbul, Turkey (2012)
6. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The europeana data model (edm). In: World Library and Information Congress: 76th IFLA General Conference and Assembly. Gothenburg, Sweden (2010)
7. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The Europeana Data Model (EDM). In: World Library and Information Congress: 76th IFLA general conference and assembly. pp. 10–15 (2010)
8. Drummond, C.: Replicability is not reproducibility: nor is it good science. In: Workshop on Evaluation Methods for Machine Learning IV (2009)
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)
10. Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N.: Offspring from reproduction problems: What replication failure teaches us. In: Proceedings of the 51st ACL. Sofia, Bulgaria (2013)
11. Garijo, D., Gil, Y.: The p-plan ontology (2013), http://www.opmw.org/model/p-plan/
12. Groth, P., Gil, Y., Cheney, J., Miles, S.: Requirements for provenance on the web. International Journal of Digital Curation 7(1) (2012)
13. Lagoze, C., van de Sompel, H.: Open archives initiative object re-use & exchange (2007), http://www.openarchives.org/ore/documents/ore-jcdl2007.pdf
14. Lagoze, C., Van de Sompel, H., Nelson, M.L., Warner, S., Sanderson, R., Johnston, P.: Object re-use & exchange: A resource-centric approach. Tech. rep. (2008)
15. McCallum, A.K.: MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu (2002)
16. McCarthy, D., Koeling, R., Weeds, J., Carroll, J.: Finding predominant word senses in untagged text. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. p. 279. Association for Computational Linguistics (2004)
17. Moreau, L., Missier, P., Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C.: PROV-DM: The PROV Data Model. Tech. rep. (2012), http://www.w3.org/TR/prov-dm/
18. Neylon, C., Aerts, J., Brown, C.T., Coles, S.J., Hatton, L., Lemire, D., Millman, K.J., Murray-Rust, P., Perez, F., Saunders, N., Shah, N., Smith, A., Varoquaux, G., Willighagen, E.: Changing computational research. the challenges ahead. Source Code for Biology and Medicine 7(2) (2012)
19. Pedersen, T.: Empiricism is not a matter of faith. Computational Linguistics 34(3), 465–470 (2008)
20. P.Vossen, Bosma, W., Rigau, G., Agirre, E., Soria, A., Aliprandi, C., de Jonge, J., Hielkema, F., Monachini, M., Bartolini, R., Frontini, F.: Kyotocore: integrated system for knowledge mining from text (2011)
21. Vanschoren, J., Blockeel, H., Pfahringer, B., Holmes, G.: Experiment databases. Machine Learning 87(2), 127–158 (2012)
22. Zaagsma, G.: Doing history in the digital age: history as a hybrid practice (2013), http://gerbenzaagsma.org/blog/16-03-2013/doing-history-digital-age-history-hybrid-practice

BiographyNet

Extracting relations between people and events

BiographyNet is a multidisciplinary project that combines expertise from history, computer science and computational linguistics. The project is a collaboration between the Netherlands eScience Center, Huygens ING and VU University Amsterdam.

This schema belongs to, and is describes in:

Ockeloen, N., Fokkens, A., ter Braake, S., Vossen, P., de Boer, V., Schreiber, A.T., Legêne, S: BiographyNet: Managing Provenance at multiple levels and from different perspectives
Workshop on Linkend Science (LISC), International Semantic Web Conference (2013)
The Network Institute, VU University Amsterdam

BIOGRAPHIES

BiographyNet uses data from the Biography Portal of the Netherlands (BP), which contains approximately 125,000 biographies from a variety of Dutch biographical dictionaries.

SEMANTICS

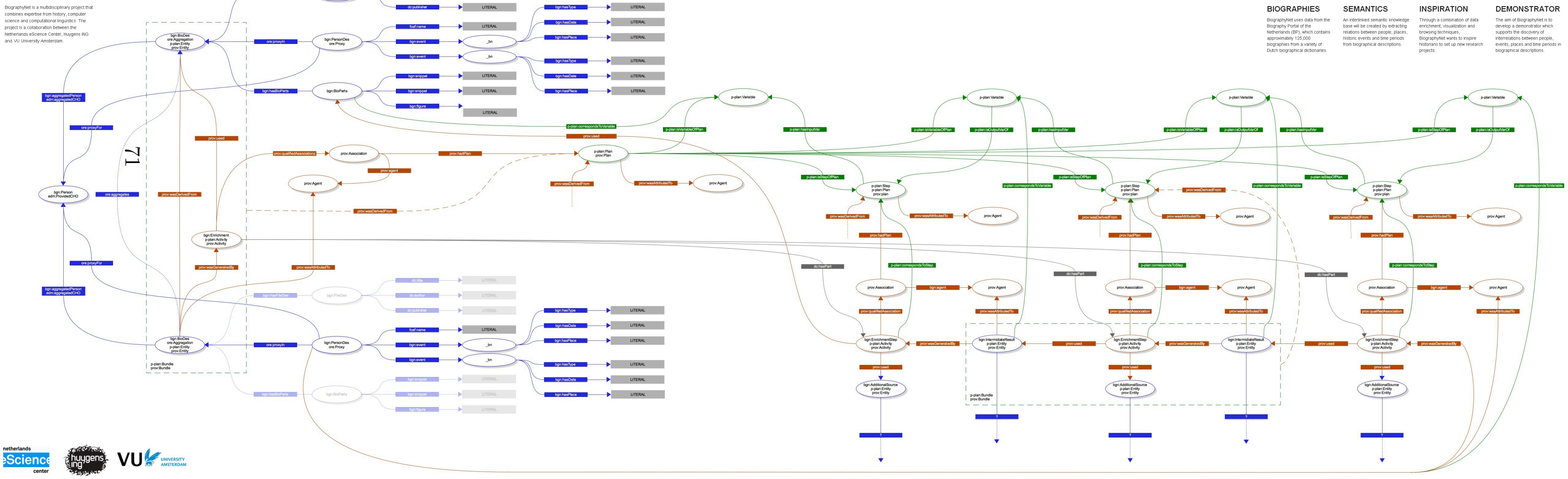
An interlinked semantic knowledge base will be created by extracting relations between people, places, historic events and time periods from biographical descriptions.

INSPIRATION

Through a combination of data enrichment, visualization and browsing techniques, BiographyNet wants to inspire historians to set up new research projects.

DEMONSTRATOR

The aim of BiographyNet is to develop a demonstrator which supports the discovery of interrelations between people, events, places and time periods in biographical descriptions.



netherlands eScience center
huygens ing
VU UNIVERSITY AMSTERDAM

Using Semantic Web Technologies to Reproduce a Pharmacovigilance Case Study

Michiel Hildebrand[1,2], Rinke Hoekstra[1] and Jacco van Ossenbruggen[1,2]

[1] VU University Amsterdam,The Netherlands
[2] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Abstract. We provide a detailed report of a reproduction study of a paper published in the International Journal of Medical Sciences (IJMS). We first use the PROV-O ontology to model our reconstruction of the computational workflow of the original experiment and to systematically explicate all information that is needed for an reproduction study. We then identify which part of the required information is published in the IJMS paper and what part is missing. We then discuss our reproduction of this workflow, following the original as much as possible. Again, we use PROV-O to precisely define our version of the workflow, including our version of the information that was missing in the IJMS paper of the study. Finally, we generalize from the specific cased described in the original paper by providing a web service that allows mining for arbitrary drug-adverse event pairs.

1 Introduction

Reproducing scientific results is often more an art than science. By describing a concrete case study we show how we used PROV-O to systematically analyse a paper from a different field, written by authors we do not personally know. We attempt to reconstruct the provenance graph of the original experiment by carefully studying the description of the method, the statistics and the results provided either directly in the paper or other sources that the paper refers to. We formalized our reconstruction using the PROV-O ontology. The formalization makes the dependencies between the intermediate steps explicit, which should allow us to systematically investigate how the results presented in the paper were computed. To reproduce the results we need to understand the input and output behavior of the computations modeled by the `prov:Activity` nodes. The properties of the input and output `prov:Entity` can help to verify wether this understanding is correct.

The paper we selected is the Open Access article *Adverse Event Profiles of 5-Fluorouracil and Capecitabine: Data Mining of the Public Version of the FDA Adverse Event Reporting System, AERS, and Reproducibility of Clinical Observations* published in the International Journal of Medical Sciences (IJMS) [12]. The paper describes a computational data mining study on public data and appears to be a good candidate for a reproduction study. We use this paper as a case study to provide insights in the problem of reproducing scientific results, we do not aim to criticize this particular paper in any way.

The topic of the paper is an example of *pharmacovigilance* which is defined by the World Health Organization as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other medicine-related problem"[3]. Computational studies play an important role in Pharmacovigilance to detect drug side-effects. Such studies are an economic way to generate hypotheses before performing costly clinical reviewing [13]. The IJMS paper [12] follows a typical scenario in pharmacovigilance: the use of a database with reports of adverse events (AE) to find disproportional correlations between a drug and an adverse reaction. In this example, the Adverse Event Reporting System of the US Food and Drug Administration (FAERS) is used to compare adverse effects of drugs.

While the FAERS database itself is publicly available, it is not trivial to reproduce the results of the experiments that use this database. Results and tools are described in scientific publications, but tools and (intermediate) results are typically not available. Our case study demonstrates in detail what prevents reproduction. From the observations of this study we derive initial requirements to support studies of drug side-effects that can be fully reproduced.

2 Related work

This section gives a brief overview of related work on data publication, scientific workflows and provenance.

The requirement for reproducibility [14] has been a key motivator for an increased interest in data sharing and publication, especially in fields dealing traditionally with ever growing datasets, e.g. [1]. Even though data sharing does not always immediately benefit the individual researcher, the potential for the scientific community is significant [15]. Funding agencies, keen on maximizing impact and reducing fraud, are now actively requiring data sharing. For example, both the US National Science Foundation and the EU now require data management plans for all proposals they consider.[4] Note that also in areas that focus on human action, such as in human computer interaction, replication has part of the research agenda[5].

However, as becomes clear in this paper as well, raw data publication (such as FAERS) is in itself not sufficient for reproducible research. Data often needs to be moulded and transformed to a new data model before it becomes suitable for answering a particular research question. This data preparation step can take between 60 to 80% of data-oriented research tasks [6]. Workflow systems [4], provide mechanisms for reproducing scientific conclusions, based on shared data.

[3] http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/

[4] See http://www.nsf.gov/bfa/dias/policy/dmp.jsp and http://europa.eu/rapid/press-release_SPEECH-13-236_en.htm

[5] http://www.cs.nott.ac.uk/~mlw/replichi.php

The benefit for individual researchers publishing a workflow, is that workflows are *executable* procedures that can be run against various inputs. Workflows can be shared and reused through social platforms such as myExperiment[6]. Curated workflow descriptions [8] combined with original data, can serve as self-contained *research objects* [3].

There are, however, two drawbacks to using a workflow system. Firstly, workflow descriptions are inevitably tied to the system used, and thus constrained to the types of operations supported by the system. Secondly, not all steps of interest in a scientific research process are necessarily of a *computational* nature, e.g. consider the information conveyed through the reuse of texts in scientific discourse. Though in its early stages, work on automatic *provenance reconstruction* [10] is a promising approach to making explicit the temporal and causal dependencies between individual elements of scientific output.

The overarching requirement for reproducible research is an explicit account of what processes and activities led from original input, albeit data, texts, other media, to the contribution of a scientific publication. The PROV standard of the W3C [11], based on ten earlier provenance models, such as the Open Provenance Model[7] and the Provenance Vocabulary[8], provides a standard vocabulary and semantics for expressing plans (workflows), process execution, dependencies between entities and processes, and agent involvement. The PROV-O ontology is a vocabulary for expressing PROV as Linked Data.[9] Most scientific workflow systems allow provenance tracking of workflow execution, and allow exporting it to PROV or a compatible format. The consumption of provenance information by applications is gradually receiving more attention [9]. The ProvBench repository[10] has the objective to bootstrap the development of systems for the visualization, analysis and understanding of provenance graphs.

3 Basic concepts in Pharmacovigilance

Various organizations maintain reporting systems of adverse events. The World Health Organization (WHO) maintains vigiBase, the US Food and Drug Administration (FDA) maintains the Adverse Event Reporting System (FAERS) and many countries maintain their own system. These organisations provide functionality for medical professionals to submit reports of adverse events that they encountered with their patients. A report in the FAERS database contains a list of the medication that the patient received and a list of adverse events. In addition it may contain information about the patient such as the gender and age. Unique of the the FAERS database is that it is publicly available on the Web. XML and CSV files for every yearly quarter starting at 2004 are available for download.

[6] See `http://www.myexperiment.org`

[7] See `http://purl.org/net/opmv/ns`

[8] See `http://purl.org/net/provenance/ns`

[9] See `http://www.w3.org/TR/prov-o/`.

[10] `https://sites.google.com/site/provbench/`

Adverse event databases are used in pharmacovigelance research to detect side effects of drugs. An important part of this research focusses on the detection of side effects of new drugs that appear on the market. The WHO has an extensive program for this research[3], and involves large scale data mining of adverse event databases. Other research focusses on the side effects of sets of specific drugs. These studies are typically motivated by clinical evidence.

Both types of research depend on methods to detect a disproportional correlation between a drug and an associated adverse event. The most common methods are the proportional reporting ratio (PRR), the reporting odds ratio (ROR), the information component (IC) and the empirical Bayes geometric mean (EBGM). All are based on the expected frequency relative to all drug event pairs that are available in the database. Calculating signals with these methods requires a 2x2 contingency table, as shown in Table 1. This table contains (a) the number of mentions of a drug together with a mention of a reaction (an adverse event), (b) the number of mentions of all other drugs and that reaction, (c) the number of mentions of the drug and all other reactions and (d) the number of mentions of all other drugs and all other reactions. According to [5] the PRR is calculated from this table using Eq. 1.

$$PRR = \frac{a/(a+c)}{b/(b+d)} \tag{1}$$

The expected value for a PRR is one and values above it indicate the strength of the association. In addition, the strength of a statistical association can be calculated using a standard chi-squared test.

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)} \tag{2}$$

According to [5] a signal is detected between a drug and an adverse event if the PRR is at least 2, the chi-squared is at least 4 and there are at least 3 or more cases mentioning the drug and the event. We refer to the literature for the details of ROR [16], IC [2] and EBGM [17]. A comparison of these methods is reported in [18].

To compute the 2x2 contingency table one needs to collect all mentions of a particular drug and a particular adverse event. Collecting the adverse event mentions is straightforward because in the FAERS database they are consistently identified with the preferred terms from the Medical Dictionary for Regulatory Activities[11] (MedDRA). The drug names in the FAERS database are, however, not standardized. The same drug may be entered in to the database in various forms. For example, drug names are entered with or without dosage information, method (e.g. oral, injection) and other additions. Some have entered the drug name, while others used the brand or trade name and again others the active ingredient. There are various spelling variations and synonyms. To properly fill the 2x2 contingency table one has to deal with the variations in drug names.

[11] `http://www.meddramsso.com/`

	Drug of interest	All other drugs	
Reaction of interest	a	b	a+b
All other reactions	c	d	c+d
	a+c	b+d	a+b+c+d

Table 1. 2x2 contingency table to calculate disproportionality measurements (adapted from [5]).

4 Case study

Our target IJMS paper [12] investigates the so called safety profiles of two types of drugs that are used in the treatment of cancer. The first drug is 5-Fluorouracil (5-FU), which was traditionally used for the treatment of solid tumors. This drug was given by injection or infusion. Due to the high risks and costs of this type of treatment the pharmaceutical industry developed a class of drugs known as oral fluoropyrimidines, from which Capecitabine is the most well known one. Clinical trials that compared the use of Capecitabine against 5-FU favor the use of the first. Due to limitations of the clinical trials the picture is, however, not complete. For example, the trials do not provide evidence for adverse events that occur at relative low frequencies. The aim of the paper is to test the conclusions drawn from the trials and provide additional evidence for lower frequency adverse events.

In the IJMS paper the authors describe the method to detect the signals for Capecitabine and 5-FU with various adverse events. As a first step towards reproduction of this study we formalized the steps and their dependencies using the PROV-O ontology. In addition, we describe the information provided in the paper that could help in the reproduction.

4.1 Provenance reconstruction

Figure 1 in Appendix A shows a reconstruction of the provenance graph for the computation of the PRR for 5-FU with the adverse event Leukopenia and the PRR of Capecitabine with Leukopenia.

Original FDA datasources The workflow starts at the bottom with datasources obtained from the FDA. The website of the FDA contains ZIP files for each yearly quarter[12]. For each quarter there are two versions available, ASCII and SGML. The former contains a dump of the database in the form of 7 CSV files, while the latter contains a single SGML file. The authors of the IJMS paper used the ASCII versions for the first quarter of 2004 up to the last quarter of 2009, a

[12] `http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm083765.htm`

total of 24 files. In the provenance graph the quarterly files are represented by individual nodes, but for the sake of clarity we do not show all nodes. The ZIP files from the FDA contain a document that describes the structure of the CSV files and instructions how to interpret them.

Report aggregation The paper mentions that the total dataset contains 2,231,029 reports. From this we conclude that an aggregation step was performed. In the provenance graph the aggregated dataset is represented by the node with the label `A.FAERS`. This aggregated dataset is the starting point for two cleanup activities. First, the authors removed superfluous reports as the data contains updated versions of a report as separate records. The paper refers to the documentation from the FDA in which it is advised to keep only the most recent report for a specific case. The resulting dataset is labeled `B.FAERS` in the provenance graph, and contains (according to the paper) 1,644,220 reports.

Drug name normalization In the second cleanup step the drug names are normalized: *all drug names were unified into generic names by a text-mining approach.* The paper does not provide details of this text-mining approach. The paper does explain that the cleanup includes the correction of spelling errors. For this purpose GNU Aspell is used to detect spelling errors and the suggested corrections are manually confirmed by *working pharmacists.* It is unclear how many spelling corrections were made. Finally, *foods, beverages, treatments (e.g. X-ray radiation), and unspecified names (e.g. beta-blockers)* were removed. It is unclear from the paper if this removal step is manual or automatic. The result of the normalization activity is represented in the provenance graph by the node `C.FAERS`.

Co-occurrence selection The paper mentions that after the drug name normalization the dataset contains 22,017,956 co-occurrences of drugs and events. A drugname and an adverse event co-occur if they are mentioned together in a report. The activity of counting co-occurrences is modeled as an explicit step and the output is the node with label `D.co-occurrences`.

Contingency table To compute the PRR values from the set of co-occurrences a 2x2 contingency table is required for each drug-adverse event pair. Populating the table requires the selection of the required subsets of co-occurrences. The graph contains activities to create the tables for 5-FU with Leukopenia and Capecitabine with Leukopenia. The resulting tables are shown as the nodes `5FU-Leukopenia` and `Capecitabine-Leukopenia`. The IMJS paper does not explicitly contain the 2x2 table for any of the drug-adverse event pairs, but using the values mentioned in the paper we can partially reconstruct the tables, see Table 2. In this table the values in bold font come from the paper, the italic ones can be trivially calculated from these. The question marks represent values which we will try to reverse engineer in the next section.

	5-FU	All other drugs	
Leukopenia	**277**	?	?
All other reactions	*40,007*	?	?
	40,284	*21,977,672*	**22,017,956**

	Capecitabine	All other drugs	
Leukopenia	**115**	?	?
All other reactions	*34,813*	?	?
	34,928	*21,983,028*	2**2,017,956**

Table 2. Partial 2x2 contingency tables for 5-FU - Leukopenia and Capecitabine - Leukopenia from the numbers provided in the IJMS paper. The numbers in italic are calculated from the numbers that are given in the paper.

PRR values The final PRR values and the results of the chi-squared test are provided in the IMJS paper. In the provenance graph they are represented as the end nodes, e.g. `PRR 5FU-Leukopenia`. Note that to recalculate the values for the PRR and chi-squared tests we need to obtain the missing values in Table 2.

5 Reproduction experiment

We first tried to recalculate the missing numbers in the 2x2 contingency tables using the information given in the paper. Next we tried to reproduce the subsets of drug-adverse event pairs that underly the 2x2 co-occurrences using the original FAERS data from the FDA website, and thus reproducing the entire workflow. Further details of the reproduction are available at the Website accompanying this paper `http://www.few.vu.nl/~michielh/lisc2013/`.

5.1 Missing numbers and formulas

Using the PRR values given in the paper and the PRR formula cited by the paper, we should be able reconstruct the missing values from the 2x2 contingency tables. Note that while we do not know values for b, the number of mentions of an adverse event in co-occurrence with all other drugs, we do know the values for $(b+d)$. Based on Eq. 1, we should thus be able to calculate the values for b by using Eq. 3.

$$b = \frac{a/(a+c)}{PRR} \times (b+d) \tag{3}$$

Knowing b, we should also be able to compute the total number of mentions of an adverse event in the database $a+b$. For example, using the PRR value

	5-FU	All other drugs	
Leukopenia	277	28,585	28,862
All other reactions	*40,007*	21,949,087	21,989,094
	40,284	*21,977,672*	22,017,956

	Capecitabine	All other drugs	
Leukopenia	115	28,747	28,862
All other reactions	*34,813*	21,954,281	21,989,094
	34,928	*21,983,028*	22,017,956

Table 3. Reproduction of 2x2 contingency tables for 5-FU - Leukopenia and Capecitabine - Leukopenia.

for 5-FU (5.282) and the numbers from the partial contingency table, Table 2, the total number of mentions of Leukopenia should be 28,887. Surprisingly, this number is different when calculated from the PRR for Capecitabine (2.520), namely 28,952. For the other adverse events mentioned in the paper we also found a difference when calculated with the PRR of 5-FU or with the PRR of Capecitabine. These differences are all bigger than can be explained by rounding errors. After more in-depth literature study we discovered that different formulas are used to calculate the PRR. For example, the IJMS paper also cites [7] that uses the formula given in Eq. 4:

$$PRR_2 = \frac{a/(a+c)}{(a+b)/(a+b+c+d)} \tag{4}$$

Unfortunately, we do not get a constant number for $a + b$ with this formula either. However, after some experimentation we discovered that with Eq. 5 we achieve a constant number for the mentions of Leukopenia, 28,862. Also for the other adverse events this formula results in a constant number. From this we conclude that while Eq. 5 is given nor cited by the IMJS paper, it is most likely the formula used to calculate all PRR values mentioned in the paper (!).

$$PRR_3 = \frac{a/c}{(a+b)/(a+b+c+d)} \tag{5}$$

Now that the total number of mentions of Leukopenia is known (a+b) we can complete the 2x2 contingency tables, see Table 3. Using this table it is also possible to, modulo rounding errors, successfully reproduce the values from the chi-squared tests with Eq. 2. Now we know how to compute the basis statistics reported by the paper, we can try to reproduce the entire experiment.

5.2 Workflow reproduction

As it is unclear how the drug name normalization was performed, we decided not to reproduce this on the entire dataset. We focus on the two drugs mentioned in the IMJS paper: 5-FU and Capecitabine. Our aim is to approximate the PRR values for these drugs and Leukopenia. The provenance graph of our reproduction is available at `http://www.few.vu.nl/~michielh/lisc2013/prov/`. We encourage the reader to access this graph. The `prov:Entity` nodes in this graph are clickable and point to the underlying data. In this way we provide access to the intermediate datasets, which is an essential ingredient to successful reproduction of computational workflows. Currently, we are investigating normalization of all drug names in the FAERS dataset.

Original FDA datasources Similarly as the study reported in the IMJS paper we downloaded the 24 quarterly dumps (from the beginning of 2004 to the end of 2009) from the FDA website.

Report aggregation by conversion to RDF We choose to aggregate the quarterly files into a single dataset by first converting them to RDF and then storing these in a triple store. The total number of reports in our RDF dataset is 2,231,038, this is 9 reports more than reported in the IMJS paper. It is unclear where the difference comes from. We can, however, confirm that the conversion to RDF did not alter the original reports, as the original CSV files combined also contain 2,231,038 unique report identifiers[13]. The conversion from CSV to RDF was performed using SWI-Prolog and the RDF conversion toolset[14]. Details of the conversion, the resulting RDF and the SPARQL endpoint are available at `http://www.few.vu.nl/~michielh/lisc2013`.

Duplicate removal The duplicate removal step was performed on the RDF dataset. We first grouped all reports with the same case number and for each group selected the report with the highest report identifier. We removed the other reports from the database. The resulting dataset contains 1,664,078 reports, this is 142 less than reported in the IMJS paper. We can't explain this difference.

Drug name normalization Instead of normalizing all the drug names, we tried to find all the mentions for our drugs of interest: 5-FU and Capecitabine. We explored four methods to find different mentions for these drug names.

1. We selected the mentions that contain the drug name itself. For Capecitabine this returns many mentions of *capecitabine*, but also many variations such as *capecitabine tablet 1000 mg*, *capecitabine roche laboratories inc* and *capecitabine 2000 mg po as divided doses daily*. In total we find 337 different mentions containing Capecitabine.

[13] The total number of unique report identifiers in the CSV files from the FDA is computed with a unix bash script: `cut -d$ -f1 DEMO0*.TXT | sort -u | wc -l`

[14] `http://semanticweb.cs.vu.nl/xmlrdf/`

2. We selected mentions of a brand name associated with the drug. For example, Capecitabine is sold under the brand name *Xeloda*. To get the brand names for a drug we used the Open Data Drug & Drug target Database, Drugbank[15]. For Capecitabine we found in Drugbank one brand name (*Xeloda*). Using this brand name 14 mentions are found in the FAERS dataset. From these mentions 4 already contain Capecitabine, e.g. *xeloda capecitabine*. For 5-FU Drugbank contains 25 brand names, such as *Adrucil* and *Fluoroplex*. For 6 of these 25 brand names, additional mentions were found in the FAERS dataset.
3. We used Drugbank to find synonyms associated with a drug name. For example, Capecitabine is also known as R340. However, no mentions of this synonym are found in the FAERS dataset. For 5-FU no synonyms are found in Drugbank.
4. We selected mentions of drug names that were spelled differently. Similarly as was reported in the IMJS paper we used GNU Aspell. Aspell contains dictionaries for many languages, but these are not very useful for drug names. Therefore, we created our own Aspell dictionary using the drug names mentioned in Drugbank. With this dictionary we generated spelling variations for all drug mentions in the FAERS dataset. For each drug mention we added the highest ranked suggestion as an alternative label to the database. For example, *capecitabine* was suggested for the drug mention *capecitabin* and *capecitapine*. When using these spelling suggestions we could retrieve for Capecitabine another 30 different mentions. From these 10 already contained the correct spelling, e.g. *capeciabine capecitabine*.

Co-occurrence selection Without drug name normalization our dataset contains a total of 23,865,029 drug-adverse event co-occurrences, 1,847,073 more than reported in the IMJS paper. This larger number of co-occurrences can be explained by the fact that we did not remove *foods, beverages, treatments (e.g. X-ray radiation), and unspecified names (e.g. beta-blockers)*, as was mentioned in the IMJS paper. In addition, drug names for a single report may contain multiple treatments each containing a different drug mention. For example, a report may contain treatment with the mention *capecitabine 500 MG* and another with the mention *capecitabine 1000 MG*. In other words, the patient received two treatments, and in the second treatment the dosage of Capecitabine was increased. Without drug name normalization these mentions are counted as two co-occurrences, whereas after normalization they will be counted as a single co-occurrence. Considering the formula for PRR in Eq. 5 this difference is reflected in the denominator, the total number of co-occurrences ($a{+}b{+}c{+}d$) as well as the total number of co-occurrences with a specific adverse event ($a{+}b$).

Contingency table Using the four methods to find drug mentions we selected the set of co-occurrences corresponding to the cells of the 2x2 contingency. The total number of co-occurrences with Leukopenia (a+b) that we found is 30,724.

[15] `http://www.drugbank.ca/`

A difference of 1,862 with the number reported in the IMJS paper. This can also be explained by the lack of drug name normalization. The total number of co-occurrences with 5-FU is 42,115, 1831 more than reported in the IMJS paper. For Capecitabine 37,973 co-occurrences are found, 3045 more than in the IMJS paper. We conclude that the four drug name selection methods find more mentions of the two drugs. Currently we are investigating if and why drug mentions are falsely included. We found 289 co-occurrences for 5-FU with Leukopenia. This is 12 more than reported in the IMJS paper. For Capecitabine 122 co-occurrences were found with Leukopenia, 7 more than the 115 reported in the IMJS paper.

PRR values Using the values in the reproduced 2x2 contingency tables and Eq. 5 the PRR for 5-FU with Leukopenia is 5.367 compared to 5.282. The chi-squared test is 1019.763 compared to 952.334. For Capecitabine with Leukopenia the PRR is 2.503 compared to 2.520 in the IMJS paper, and the chi-squared test is 109.661 compared to 103.730.

6 Discussion

Reproducing the study described in the IMJS paper required substantial effort, it was difficult to verify the results of the intermediate datasets and almost impossible to analyze the differences in the reproduction. And this is all despite the fact that the IMJS paper of the case study at first sight clearly describes the method and results. By formalizing the computational workflow in PROV-O it became possible to systematically investigate the intermediate steps. We believe that sharing such provenance graphs is a first step in simplifying the reproduction of computational workflows. The next step is to also make the content of the `prov:Entity` nodes available, and ultimately the computational processes that underly the `prov:Activity` nodes. We hope that the clickable provenance graph we made available at `http://www.few.vu.nl/~michielh/lisc2013/prov/` serves as an example.

Acknowledgments

This work was funded under the Dutch COMMIT program as part of the Data2Semantics and SEALINCMedia projects.

References

1. H. Akil, M. E. Martone, and D. C. Van Essen. Challenges and opportunities in mining neuroscience data. *Science*, 331(6018):708–712, 2011.
2. A. Bate, M. Lindquist, I. Edwards, S. Olsson, R. Orre, A. Lansner, and R. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, 54(4):315–321, 1998.

3. K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garcia-Cuesta, J. Gmez-Prez, G. Klyne, K. Page, M. Roos, J. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. Goble. A workflow-centric research objects: A first class citizen in the scholarly discourse. In *Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012)*, Heraklion, Greece, May 2012.
4. E. Deelman, D. Gannon, M. Shields, and I. Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, 25(5):528–540, 2009.
5. S. J. W. Evans, P. C. Waller, and S. Davis. Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6):483–486, 2001.
6. D. Garijo, P. Alper, K. Belhajjame, O. Corcho, Y. Gil, and C. Goble. Common motifs in scientific workflows: An empirical analysis. In *8th IEEE International Conference on eScience*, USA, 2012. IEEE Computer Society Press.
7. A. Goald. Practical pharmacovigilance analysis strategies. *Pharmacoepidemiology and drug safety*, 12(7):559–574, 2003.
8. C. Goble, R. Stevens, D. Hull, K. Wolstencroft, and R. Lopez. Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, 9(6):506–517, 2008.
9. P. Groth and J. Frew. Proceedings of the 4th international conference on provenance and annotation of data and processes. 2012.
10. P. Groth, Y. Gil, and S. Magliacane. Automatic metadata annotation through reconstructing provenance. In *Third International Workshop on the role of Semantic Web in Provenance Management, ESWC 2012*, 2012.
11. P. Groth and L. Moreau. PROV-Overview: An Overview of the PROV Family of Documents. Working group note, W3C, Apr. 2013. `http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/`. Latest version available at `http://www.w3.org/TR/prov-overview/`.
12. K. Kadoyama, I. Miki, T. Tamura, J. Brown, T. Sakaeda, and Y. Okuno. Adverse event profiles of 5-fluorouracil and capecitabine: Data mining of the public version of the fda adverse event reporting system, aers, and reproducibility of clinical observations. *International Journal of Medical Sciences*, 9(1):33–39, 2012.
13. M. Liu, M. E. Matheny, Y. Hu, and H. Xu. Data mining methodologies for pharmacovigilance. *ACM SIGKDD Explorations Newsletter*, 14(1):35–42, 2012.
14. J. P. Mesirov. Accessible reproducible research. *Science*, 327(5964):415–416, 2010.
15. H. A. Piwowar, R. S. Day, and D. B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PloS one*, 2(3):e308, Jan. 2007.
16. K. Rothman, S. Lanes, and S. Sacks. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and drug safety*, 13(8):519–523, 2004.
17. A. Szarfman, S. Machado, and R. O'Neill. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the us fda's spontaneous reports database. *Drug safety : an international journal of medical toxicology and drug experience*, 25(6):381–392, 2002.
18. E. van Puijenbroek, A. Bate, H. Leufkens, M. Lindquist, R. Orre, and A. Egberts. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, 11(1):3–10, 2002.

Appendix A

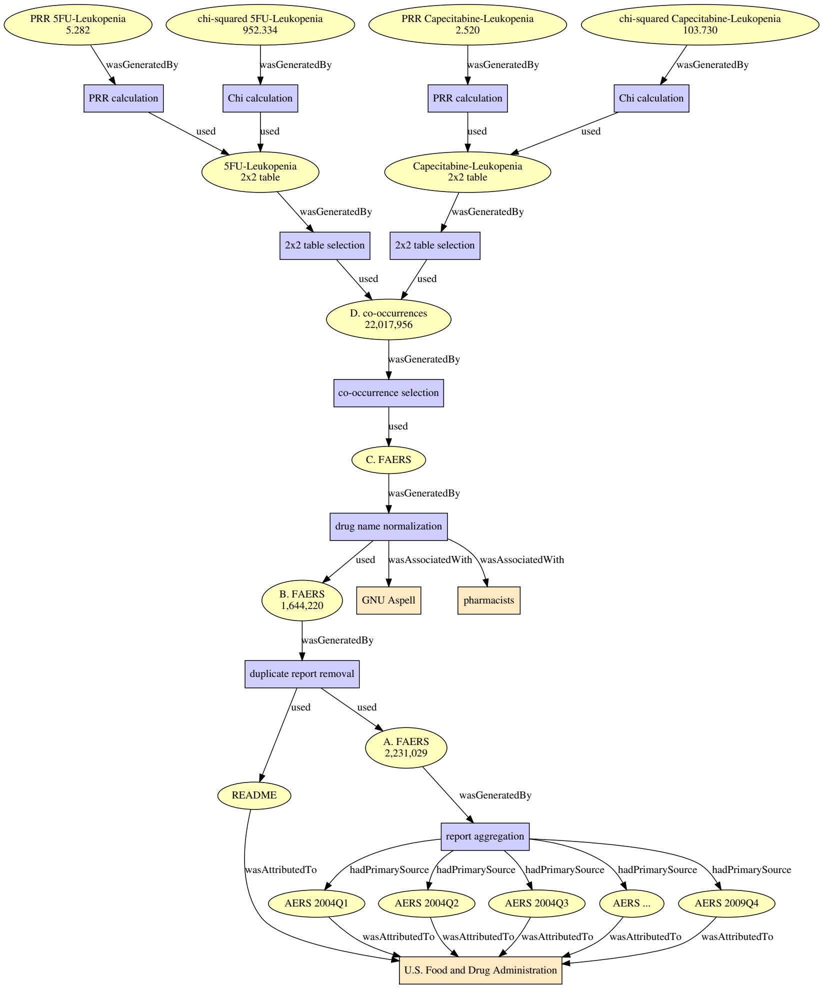


Fig. 1. Reproduction of the provenance graph corresponding to the computational workflow described in [12].