Semantic models in biomedicine: Building interoperating ontologies for biomedical data representation and processing in pharmacovigilance

by

MELANIE COURTOT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2014

Abstract

It is increasingly challenging to analyze the data produced in biomedicine, even more so when relying on manual analysis methods. My hypothesis is that using a common representation of knowledge, implemented via standard tools, and logically formalized can make those datasets computationally amenable, help with data integration from multiple sources and allow to answer complex queries. The first part of this dissertation demonstrates that ontologies can be used as common knowledge models, and details several use cases where they have been applied to existing information in the domain of biomedical investigations, clinical data and vaccine representation. In a second part, I address current issues in developing and implementing ontologies, and proposes solutions to make ontologies and the datasets they are applied to available on the Semantic Web, increasing their visibility and reuse. The last part of my thesis then builds upon the first two, and applies their results to pharmacovigilance, and specifically to analysis of reports of adverse events following immunization. I encoded existing standard clinical guidelines from the Brighton Collaboration in Web Ontology Language (OWL) in the Adverse Events Reporting Ontology (AERO) I developed within the framework of the Open Biological and Biomedical Ontologies Foundry. I show that it is possible to automate the classification of adverse events using the AERO with very high specificity (97%). I also demonstrate that AERO can be used with other types of guidelines. Finally, my pipeline relies on open and widely used data standards (Resource Description Framework (RDF), OWL, SPARQL) for implementation, making the system easily transposable to other domains. This thesis validates the usefulness of ontologies as semantic models in biomedicine enabling automated, computational processing of large datasets. It also fulfills the goal of raising awareness of semantic technologies in the clinical community of users. Following my results the Brighton Collaboration is moving towards providing a logical representation of their guidelines.

Preface

- In Chapter 3, a version of section 3.2 was published as "Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, et al. *Modeling biomedical experimental processes with OBI.* J Biomed Semantics, 1(Suppl 1):S7, 2010". I was the core developer in charge of most development in the OBI consortium at this time, and participated in development of the general framework as well as implementation of the models. I produced the release OWL file on which the manuscript is based. I participated in the implementation of all the use cases and addressing their representational needs within OBI and IAO, as core developers of both those resources. My work focused on the neuroscience investigation (Use case 1) in collaboration with Dirk Derom and Alan Ruttenberg, as well as the vaccine protection investigation (Use case 2) in collaboration with Yongqun He. I reviewed and edited the manuscript. A version of section 3.3 was published as "Yongqun He, Zuoshuang Xiang, Thomas Todd, Mélanie Courtot, RR Brinkman, Jie Zheng, Christian J Stoeckert Jr, James Malone, Philippe Rocca-Serra, Susanna-Assunta Sansone, et al. *Ontology representation and anova analysis of vaccine protection investigation.* In Bio-Ontologies 2010: Semantic Applications in Life Sciences, 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB): 2010; Boston, MA, USA. August 11, volume 13, page 4, 2010." I participated in the implementation of the use cases in VO, OBI and IAO. Yongqun He, Zuoshuang Xiang and Thomas Todd applied it to the Brucella case. I reviewed and edited the manuscript.

- Portions of Chapter 4 were prepared for submission as "Yongqun He, Zuoshuang Xiang, Lindsay Cowell, Alexander D. Diehl, Harry Mobley, Bjoern Peters, Alan Ruttenberg, Richard H. Scheuermann, Ryan R. Brinkman, Mélanie Courtot, Chris Mungall, Fang Chen, Thomas Todd, Lesley Colby, Howard Rush, Trish Whetzel, Mark A. Musen, Brian D. Athey, Gilbert S. Omenn, Barry Smith *VO: Vaccine Ontology*". I participated in the development of the resources, developers discussions, calls and meetings, as well as manuscript preparation and editing. I was a core developer of the Vaccine Ontology (VO), and participated actively in establishing the original framework in terms of classes and relations. I directly contributed to all terms described in this chapter, amongst others. I formalized knowledge for Canadian vaccines, while my collaborators added US ones. Edits to the OWL file were done by Yongqun Oliver He following our discussions. VIOLIN and literature-based mining were done at University of Michigan. Permission to reproduce parts of this paper for the purpose of this thesis was obtained from all co-authors.

- A version of Chapter 5 was published as "Philippe Rocca-Serra, Alan Ruttenberg, Martin J O'Connor, Patricia L Whetzel, Daniel Schober, Jay Greenbaum, Mélanie Courtot, Ryan R Brinkman, Susanna Assunta Sansone, Richard Scheuermann, et al. *Overcoming the ontology enrichment bottleneck with quick term templates.* Applied Ontology, 6(1):13-22, 2011", and is reprinted with permission from IOS Press. I was core developer of the OBI consortium and extensively contributed to all aspects of development, including Quick Term Template. Philippe Rocca-Serra led this work, to which I contributed with 6 other authors. All authors

(11 + the consortium) participated to the manuscript preparation.

- In Chapter 6, a version of section 6.2 was published as "Mélanie Courtot, Chris Mungall, Ryan R. Brinkman, and Alan Ruttenberg. *Building the OBO Foundry - one policy at a time.* In Proceedings of the International Conference on Biomedical Ontology (ICBO2011), 2011". I worked in collaboration with Alan Ruttenberg and Chris Mungall on devising and implementing the policies described. I wrote the original draft of the ID specification, the documentation for the common metadata scheme and was the lead developer of the MIREOT. I drafted the original manuscript. A version of section 6.3 was published as "M.Courtot, F.Gibson, A.L.Lister, J.Malone, D.Schober, R.R.Brinkman and A.Ruttenberg. *MIREOT: The minimum information to reference an external ontology term.* Applied Ontology, 6(1):23-33, 2011", and is reprinted with permission from IOS Press. In collaboration with Alan Ruttenberg, I articulated the problems, devised the guidelines supporting the methodology and provided an implementation of the specification. I drafted the original manuscript. A version of section 6.4 was published as "Zuoshuang Xiang, Mélanie Courtot, Ryan R Brinkman, Alan Ruttenberg and Yongqun He". *Ontofox: web-based support for ontology reuse.* BMC research notes, 3(1):175, 2010. Ontofox implements the MIREOT mechanism I developed. I participated in the system development via initial prototype development, discussion, testing, feedback and suggestions. I contributed extensively to editing of the original draft manuscript. Zuoshuang Xiang was in charge of the server implementation and maintenance.

- A version of Chapter 7 was prepared for submission for peer-review publication as "Zuoshuang Xiang, Mélanie Courtot, Chris Mungall, Alan Ruttenberg, and Yongqun He. *Ontobee: A Linked Data Server for OWL Ontology Terms*". Ontobee implements the dereferencing prototype mechanism I developed with Alan Ruttenberg. I identified issues, reviewed existing work and developed the original prototype for publication of OBO ontologies on the Semantic Web (Linked Ontology Data) on which Ontobee is based. I participated in Ontobee's development via discussion, testing, feedback and suggestions. I contributed extensively to editing of the original draft manuscript. Zuoshuang Xiang was in charge of the server implementation and maintenance. Figures were produced by Yongqun He, with the exception of Figure 7.6 which I made with Alan Ruttenberg.

- Parts of Chapter 8 were published as "Mélanie Courtot, Jie Zheng, Chris Stoeckert, Ryan Brinkman and Alan Ruttenberg *Diagnostic criteria and clinical guidelines standardization to automate case classification* Proceedings of the International Conference on Biomedical Ontologies (ICBO) 2013." and "M. Courtot, R. R. Brinkman, and A. Ruttenberg. *The logic of surveillance guidelines: an analysis of vaccine adverse event reports from an ontological perspective.*" In both cases I performed the ontology development and drafted the manuscript. Jie Zheng implemented the Malaria use case described in Section 8.5.

- A version of Chapter 9 was accepted by PLoS ONE on February 25th as "Mélanie Courtot, Ryan R. Brinkman, and Alan Ruttenberg. *The logic of surveillance guidelines: An analysis of vaccine adverse event reports from an ontological perspective*". I collected the datasets, performed the experiments, analyzed the data and wrote the manuscript draft.

- No ethics approval was required for this research, as confirmed by the UBC BCCA Research Ethics Board and supported by article 2.4 of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans document [1] which states that "REB review is not required for research that relies exclusively on secondary use of anonymous information, or

anonymous human biological materials, so long as the process of data linkage or recording or dissemination of results does not generate identifiable information.".

Table of Contents

Appendices

List of Tables

List of Figures

Abbreviations

ABC Automatic Brighton Classification
AEFI Adverse Event Following Immunization
AEO Adverse Events Ontology
AERO Adverse Events Reporting Ontology
AERS Adverse Event Reporting System
AUC Area Under the Curve
BCCD Brighton Collaboration Case Definition
BCPNN Bayesian Confidence Propagation Neural Network
BFO Basic Formal Ontology
CAEFI Canadian Adverse Event Following Immunization
CAEFISS Canadian Adverse Events Following Immunization Surveillance System
CARO Common Anatomy Reference Ontology
CDC Center for Disease Control and Prevention
CFEP Canadian Field Epidemiology Program
ChEBI Chemical Entities of Biological Interest
CL Cell Type
DC Dublin Core
EBS Empiric Bayesian Screening
EDC Electronic Data Capture
EHR Electronic Health Record
FDA US Food and Drug Administration
FMA Foundational Model of Anatomy
FOIA Freedom of Information Act
GO Gene Ontology
GPS Gamma Poisson shrinkage
IAO Information Artifact Ontology
IC Information component
IDO Infectious Disease Ontology
ILI Influenza-like Illness
IMPACT Immunization Monitoring Program ACTive
InfluenzO Influenza Ontology
IR Information Retrieval
IRI Internationalized Resource Identifier
LOD Linked Open Data
MedDRA Medical Dictionary of Regulatory Activities
MGPS Multi-gamma Poisson shrinker
MIREOT Minimum Information to Reference an External Ontology Term
MMR Measles, Mumps, and Rubella
MP Mammalian Phenotype Ontology
MSSO Maintenance and Support Services Organization

NCBO National Centre for Biomedical Ontology
NEMO Neural ElectroMagnetic Ontologies
NIAID/FAAN National Institute of Allergy and Infectious Diseases/Food Allergy and Anaphylaxis Network
NIF Neuroscience Information Framework
NLP Natural Language Processing
OAE Ontology of Adverse Events
OBI Ontology for Biomedical Investigations
OBO Open Biomedical Ontologies
OGMS Ontology for General Medical Science
OMRE Ontology of Medically Relevant Entities
OWL Web Ontology Language
PATO Phenotypic Quality Ontology
PCIRN PHAC/CIHR Influenza Research Network
PHAC Public Health Agency of Canada
PRO Protein Ontology
PRR Proportional Reporting Ratios
PHSA Provincial Health Services Authority
QTT Quick Term Template
RDF Resource Description Framework
RDFS Resource Description Framework (RDF) Schema
RIF Rule Interchange Format
RO Relations Ontology
ROC Receiver Operating Characteristic
ROR Reporting Odds Ratios
RR Relative Risk
RRR Relative Report Rate
SKOS Simple Knowledge Organization System
SMQ Standardised MedDRA Query
SNOMED-CT Systematized NOmenclature of MEDicine Clinical Terms
SPARQL SPARQL Protocol and RDF Query Language
SO Sequence Ontology
UO Unit Ontology
URI Uniform Resource Identifier
VAERS Vaccine Adverse Event Reporting System
VENICE Vaccine European New Integrated Collaboration Effort
VO Vaccine Ontology
WHO World Health Organization

Acknowledgements

I would like to thank my supervisor, Dr. Ryan Brinkman, who encouraged me to start graduate studies after having worked as an engineer in his lab. I am extremely grateful for his continuous help and support, both personally and professionally.

This thesis would not had been possible without contributions from my thesis committee, Drs Paul Pavlidis, Margaret-Anne Storey and Raymond Ng, and members of my defence examining committee: Dr Haydn Pritchard (chair), Drs Julie Bettinger and Rachel Pottinger (university examiners) and Dr. Pascal Hitzler (external reviewer). Thanks to Dr. Mark Wilkinson for his support in multiple occasions, and his participation in the committee at initial stages.

I have been immensely blessed to have the opportunity to work with amazingly talented individuals, who all spent time discussing and explaining their area of expertise - thanks to all my collaborators. I am especially thankful to Alan Ruttenberg for his ongoing advice and collaboration, helping me improve my projects on multiple occasions, and introducing me to other areas and types of problems in the general field of data sharing and query answering. I also wish to acknowledge my colleagues Drs Jie Zheng, James Malone, Bjoern Peters, Christian Stoeckert, William Bug, Chris Mungall, Barry Smith, Albert Goldfain, Richard Scheuermann and Lindsey Cowell for their collaboration on various ontology development projects. Drs Robert Pless, Jan Bonhoeffer, Barbara Law, Jean-Paul Collet and Ms Julie Laflèche helped me understand vaccine safety aspects and supported my work in multiple occasions.

Thanks to Dr Nicolas Le Novère for giving me the first opportunity to work on knowledge representation and develop the Systems Biology Ontology, as well as my other past supervisors, Drs Franc Pattus, Renaud Wagner and Christos Ouzounis. You all showed me what research could be like and inspired me to take the leap and go back to school.

This thesis was partly supported by funding from the Public Health Agency of Canada/Canadian Institutes of Health Research Influenza Research Network (PCIRN), and the Michael Smith Foundation for Health Research.

Finally, thanks to my family for their unrelenting support: my parents, Jean-Claude and Claudine, my sister Julie, my partner Brian and my daughter Hannah.

To Hannah

Chapter 1

Overview

Assessment of pharmacovigilance data is a largely manual, time-consuming process [2]. Additionally, analysis of large datasets, such as those in current reporting systems, can be challenging [3]. As a result, rapid detection of safety issues can be hampered by the methods used for surveillance, even more so when a large volume of data such as in the 2009-2010 H1N1 pandemic is being collected. In that context, I hypothesized that ontologies and Semantic Web technologies can be used to make biomedical research in general, and pharmacovigilance in particular, more accurate and reproducible.

1.1 Research questions

This dissertation is divided into three major sections. The first section aims at providing a means of representing knowledge, specifically in the biomedical domain, using ontologies and the Web Ontology Language (OWL). The second section introduces some of the issues raised by developing multiple resources aimed at working together in supporting multiple applications, across a large consortium of ontology developers, the OBO Foundry. Finally, the third section relies on the other two and applies their findings to the domain of pharmacovigilance, leading to the development of the AERO and its application to automated classification of adverse events.

Specifically, I investigate and answer the following research questions:

1. Can ontologies be used to encode biomedical knowledge, and specifically biomedical investigations and pharmacovigilance, in a standard, unambiguous way, allowing semantic querying (i.e., be complex enough to encode all logical aspects while maintaining reasoning capabilities)?

    - Can biomedical investigations and pharmacovigilance be accurately represented using ontologies? I hypothesized that a standard way of modeling information would improve description of experimental processes, hence data comparison and integration.

2. What are the elements required for supporting large consortiums of ontology developers building compatible resources for publication on the Semantic Web?

    - How can a suitable framework for development of collaborative, interoperating resources be provided?
    - What are some of the issues encountered when working with those large interoperating biomedical resources, and how they be overcome?

3. Can adverse event classification in pharmacovigilance be improved through the use of ontologies to automate the process?

    - Can the logic of a pharmacovigilance clinical guideline be encoded as an ontology? Does a standard and logically formalized representation of the Brighton Collaboration case definitions enhance data quality and allow for automatic processing of adverse events reports? I hypothesized that a standard and logically formalized representation of the Brighton Collaboration case definitions would enhance data quality and allow for automatic processing of adverse events reports.
    - Will establishing a mapping between this ontology and another resource (terminology, other ontology) used to annotate existing AE reports datasets allow us to infer that the data is of the type of a specific ontology class (i.e., derive a diagnosis according to the selected guideline)? I hypothesized that using the AERO and a custom mapping, adverse event reports could be automatically classified according to a Brighton case definition.
    - What is the efficiency of this classification? I hypothesized that the classification would be more efficient in terms of time and cost than performed by human review.

1.2 Contributions and impact

1. The first part of my thesis details how I solved representation issues in the biomedical domain in areas that formed the basis of my later work. I actively contributed to the development of seven ontologies addressing different kind of problems and domains:

    - The Ontology for Biomedical Investigations (OBI), which models investigations, including their plans and objectives, their realization by experimental processes, as well as participants involved,

- The Information Artifact Ontology (IAO), which addresses the need for representation of data and information entities, such as data item, directive information entities (including guidelines), e-records etc.,
- The Basic Formal Ontology (BFO), an upper-level ontology supporting analysis and integration,
- The Vaccine Ontology (VO), which focus is on representation of vaccination and associated immunologic responses, as well as vaccines and vaccine components,
- The Infectious Disease Ontology (IDO) and the Ontology for General Medical Science (OGMS), which aim at representing infectious diseases and clinical data,
- The Adverse Events Reporting Ontology (AERO), an ontology representing guidelines used in pharmacovigilance.

I was in each case part of the core developers group and contributed significantly to building the resources, either general framework such as critical terms and relations between them, or specific such as representation of clinical guidelines in the context of the Ontology for General Medical Science (OGMS). I brought a pharmacovigilance perspective to this work, and generally worked on representation that I expected would contribute to my research goal. Representation work culminated with me creating the AERO.

Chapter 3 describes how biomedical investigations can be modeled in a standard, unambiguous way which allows semantic querying. It details how some representation issues in the biomedical domain in areas, that formed the basis of following chapters, were solved. Specifically, it presents three use cases that were modeled within the Ontology of Biomedical Investigations (OBI). I participated in the implementation of all the use cases and addressing their representational needs within OBI and IAO, as core developers of both those resources. My work focused on the neuroscience investigation (Use case 1) in collaboration with Dirk Derom and Alan Ruttenberg, as well as the vaccine protection investigation (Use case 2) in collaboration with Yongqun He. Larissa Soldatova was the main developer of Use case 3.

Chapter 4 introduces the Vaccine Ontology (VO), another resource related to my work towards pharmacovigilance data representation. Amongst others, details of the vaccination process and vaccine composition can be captured via the VO. I participated in the implementation of the use case in VO, OBI and IAO, while my collaborators at the University of Michigan applied it to the Brucella case.

Having a common, standardized representation of biomedical knowledge will improve the ability of exchanging and integrating data, with the goal of answering complex queries across multiple data sources.

2. The contents of an ontology is only part of what is necessary for adoption. The second part of my thesis concerns how to support development and dissemination of resources, such as ontologies, that are collaboratively constructed within a consortium of biomedical specialists. I used an emerging technology, the Semantic Web, as a publication medium developing methods and practices enabling dissemination of these resources using Semantic Web technologies.

    - I developed the MIREOT to make it feasible to work with parts of other ontologies, particularly when tools such as editors and reasoners could not effectively work with full versions of those ontologies.
    - I co-designed the OntoFox, a web-server implementing the MIREOT mechanism through an accessible web interface.
    - With one collaborator I developed the original prototype for publication of OBO ontologies on the Semantic Web (Linked Ontology Data). OBO format ontologies contain many essential terms which were previously not as easily usable, and which were unavailable for use in the Semantic Web.
    - I was one of the designers of Ontobee, a server implementing this prototype, and which is now the default server for terms from all OBO ontologies.
    - I was one of the designers of Quick Term Template (QTT), which makes it easier for scientists to define many ontology classes whose definitions follow a common pattern by using common spreadsheet applications.

Chapter 5 details the rationale, design and implementation for the Quick Term Templates (QTT) tool which allows semi-automated addition of multiple OWL classes in an ontology when those classes all adhere to the same design pattern.

Chapter 6 concerns how to support development and dissemination of resources, such as ontologies, that are collaboratively constructed within a consortium of biomedical specialists. It explores how large biomedical resources can be made practically (re) usable to the community of ontologies developers. The MIREOT mechanism allowing reuse of terms from

external resources, as well as its implementation within the OntoFox server are described, in Sections 6.3 and 6.4 respectively.

An emerging technology, the Semantic Web, is used as a publication medium developing methods and practices that enable dissemination of these resources using Semantic Web technologies. **Chapter 7** details how, after resources are built using the mechanisms detailed in earlier chapters, they can be published on the Semantic Web. The Ontobee server was developed to provide a human-friendly HTML interface as well as RDF for consumption by machines.

By enabling ontology building using Semantic Web technologies, my work helps fulfill goals of both communities, towards improving understanding of data semantics by machines.

3. In the third part of my thesis I describe my implementation of a system for adverse event classification in pharmacovigilance based on the approaches I developed in my earlier work. Specifically,

   - I created the Adverse Events Reporting Ontology (AERO) with the goal of creating a more rigorous encoding of guidelines about Adverse Events Following Immunization (AEFIs), with the Brighton guidelines for Anaphylaxis forming the nucleus of this effort.
   - I collected several adverse events datasets, and translated them into a Brighton-annotated format
   - I exported them as OWL documents represented using AERO and classified the adverse events using recently developed reasoners - a central component of Semantic Web technology
   - I validated my classification results against existing tools/gold standards
   - I developed a screening algorithm that is more efficient than those previously published

   Chapter 8 details the development of the Adverse Event Reporting Ontology (AERO) and how it allows logical translation of clinical guidelines used in pharmacovigilance, such as the Brighton Anaphylaxis guideline.

   Chapter 9 describes the implementation of a system for adverse event classification in pharmacovigilance based on the approaches developed in earlier chapters. Specifically, several adverse events datasets were collected and translated into a Brighton-annotated format, then

exported as OWL (Web Ontology Language) documents represented using AERO. Adverse events were then classified using recently developed reasoners - a central component of Semantic Web technology. Classification results were evaluated against existing tools/gold standards, and a screening algorithm that is more efficient than those previously published was developed.

Based on experience developing the system, and in collaboration with the Brighton Collaboration and the Public Health Agency of Canada (PHAC), ways to improve AEFI reporting standards and systems are proposed. This last part of the thesis exemplifies how practically, in clinical settings and with a real-world example of importance to public health, semantic resources can help improve processing of the ever-increasing collected data.

Chapter 2

Background

Pharmacovigilance focuses on safety of medicinal products, with the specific tasks of collecting, detecting, assessing, monitoring and preventing adverse events they may cause [4].

The thalidomide disaster of the 1960s [5] had profound impact on drug safety assessment and regulatory aspects [6], and the WHO international monitoring of drug safety was established shortly thereafter. As of the end 2010, 134 countries were part of the WHO pharmacovigilance program[1]. While all practitioners agree on the importance of reporting adverse events in increasing public health safety, current methods used for spontaneous adverse events reporting are not sufficient, mitigating their usefulness.

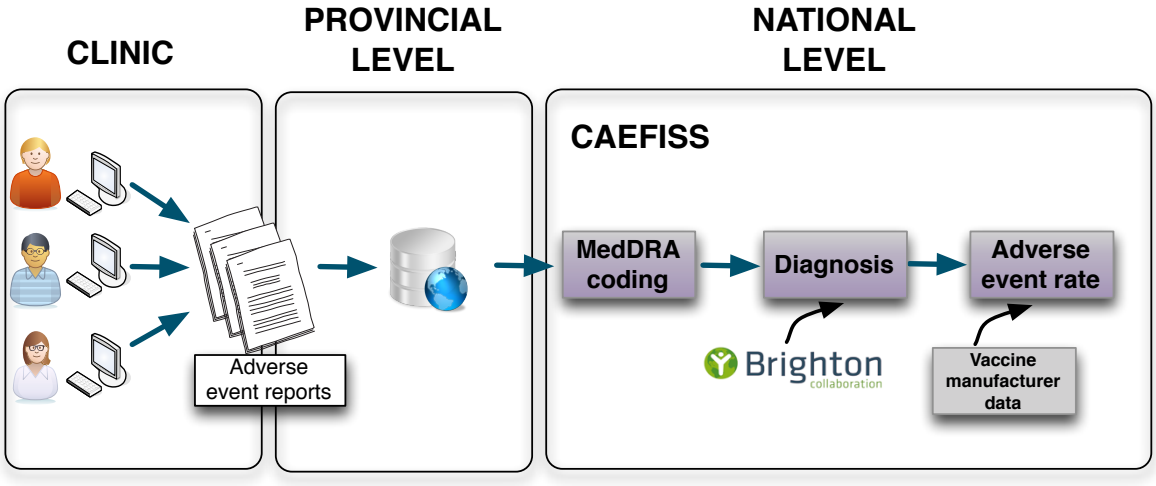


Figure 2.1: The adverse event reporting pipeline in Canada. Reports are entered at the clinic level, and forwarded to the provincial health agency. Reports are aggregated and sent to the CAEFISS database at the national level, where MedDRA codes are added and medical officers try and assess or confirm the diagnosis. Finally, based on information such as number of doses manufactured, an adverse event rate is estimated.

In Europe, the Vaccine European New Integrated Collaboration Effort (VENICE) [7] group

[1] http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/index.html

reports [8] that only 71% (17/24) of the countries states have adopted a classification of AEFIs, and that those chosen classifications are heterogeneous: 38% WHO[2] and 62% other or not specified. In the US, monitoring is done via the Adverse Event Reporting System (AERS) [9] and Vaccine Adverse Event Reporting System (VAERS) [10] systems for drugs and vaccines respectively. In Canada, the Public Health Agency of Canada (PHAC) administers the Canadian Adverse Events Following Immunization Surveillance System (CAEFISS). In both countries, systems aggregate data at a national level and rely on the Medical Dictionary of Regulatory Activities (MedDRA) to encode adverse events. Several studies highlight the potential issues in using MedDRA for adverse event reporting, ranging from inaccurate reporting as several terms are non-exact synonyms, to lack of semantic grouping features impairing processing in pharmacovigilance [11, 12, 13, 14]. Additionally, in many systems only the adverse event code as determined by the system (e.g., resulting from parsing the textual input) is saved, and information about signs and symptoms used in the determination of that code are lost. This limits the ability of analysts to review the set of symptoms observed in order to establish a consistent diagnosis. Finally, this code is not linked to any definition. This in turn may lead to heterogeneity in the diagnoses recorded [15] - physicians may have slightly different interpretations of what constitutes an anaphylactic reaction for example, as shown on Figure 2.3.

The resultant lack of consistency limits the ability to query and assess important safety issues the resulting datasets might otherwise support.

2.1 Adverse Events Following Immunization (AEFIs)

2.1.1 What is an adverse event?

The Guidance for Clinical Safety Data Management: Definitions and Standards for Expedited Reporting [16], defines an adverse event as "Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment." The guide then adds "An adverse event (AE) can therefore be any unfavourable and unintended sign (including an abnormal laboratory finding, for

[2]The reports can be difficult to even interpret - the WHO Adverse Reaction Terminology (WHO-ART) is a non-open terminology and only a 1997 version appears to be publicly visible, hosted at `http://bioportal.bioontology.org/ontologies/40404`. It lacks many terms that are essential for AE reporting, such as those related to seizure. More importantly, WHO-ART follows a 4-level structure similar to MedDRA, and therefore suffers some of the same defects.

example), symptom, or disease temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product." The Report of Adverse Event Following Immunization (AEFI) user guide [17] from the Public Health Agency of Canada (PHAC) adheres to this definition and adapts it for AEFI reporting: "An AEFI is any untoward medical occurrence in a vaccine which follows immunization and which does not necessarily have a causal relationship with the administration of the vaccine". Not detailed in this statement is the additional fact that reporting guidelines often provide protocols for determining and reporting the likelihood that specific pathological processes have occurred, and that such protocols and reporting conventions differ from jurisdiction to jurisdiction, from investigation to investigation, and by symptom and severity. Therefore adverse events as recorded in reports, contrary to what might otherwise be presupposed, are not necessarily processes, are not necessarily of the type reports say they are, are not necessarily causally related to the intervention which led to them being reported, and the terms used to describe them are not necessarily univocal. In particular, adverse event is distinguished from adverse side effect, where the latter is of a type determined to be causally related to the intervention. This matches the usage made for example within the VAERS [10] that mentions "VAERS collects data on any adverse event following vaccination, be it coincidental or truly caused by a vaccine".

2.1.2 What may cause an adverse event?

Different etiologies are at play in terms in possibly causing adverse events. The most obvious cause is the vaccination itself: either in terms of poor injection technique or stress generated by the process, which may in turn result in vaso-vagal type of events, including fainting or hypotonic/hyporesponsive episodes. Components of the vaccine themselves may also cause diverse reactions. For example, the Bacille Calmette-Guérin (BCG) vaccine has been shown to cause local swelling of the lymph nodes (suppurative lymphadenitis [18, 19]), even more so when more virulent strains were used for vaccination (such as with the vaccine BCG-Pasteur Intradermal P, Charge R 5520 [20]). The host immune response plays a major role in the occurrence of adverse event, as described in Table 2.1 The most common one is probably local inflammation due to the innate immune response, which results in redness and swelling at injection site. The Arthus reaction [21], an hypersensitivity type III reaction, similarly causes redness and swelling, but with severe associated pain - it is linked to antigen deposit meeting high quantities of antibody in presensitized patients which already had circulating antibody. Systemic inflammatory responses such as fever, irritability, nausea, vomiting of general muscle aches can also occur; their etiology is less clear,

Table 2.1: Potential immune-mediated reactions to vaccines. Adapted from [24]

Immune mediated reaction	Frequent clinical manifestation
IgE mediated	Urticaria, angioedema, rhinoconjunctivitis, bronchospasm, anaphylaxis, gastrointestinal disorders (diarrhea, abdominal cramping, vomiting)
Immune complex (IgG)	Vasculitis, myocarditis
T-cell mediated	Maculopapular exanthema, eczema, acute generalised exanthematous pustulosis (AGEP), erythema multiforme
Non-IgE mediated (pseudo-allergic)	Urticaria, angioedema, anaphylactoid reactions, gastrointestinal disorders
Autoimmune, inflammatory	Thrombocytopenia, vasculitis, polyradiculoneuritis, macrophagic myofasciitis, rheumatoid arthritis, Reiter's syndrome, sarcoidosis (juvenile), bullous pemphigoid, lichen planus, Guillain-Barré syndrome, polymalgia

though host factors (e.g., age, gender, genetics) seem to play a role in susceptibility. Type I hypersensitivity reactions include urticaria, angio-edema and anaphylaxis - the latter being used as case study throughout this thesis. As is shown in Table 2.1, it can be hard to distinguish between anaphylactoid reactions (non-IgE mediated, row 4) and true anaphylaxis reactions (IgE mediated, row 1). For example, a new type of adverse event, the oculo-respiratory syndrome (ORS) was identified in Canada in 2000 [22], and only skin testing [23] showed that it was not a type I (i.e., IgE mediated) hypersensitivity reaction.

2.1.3 When are adverse events reported?

Current guidelines [17] specify that events should be reported on the basis of their temporal association with the medical intervention. For example, in the case of AEFIs depending on (i) the type of immunizing agent (30 days after live vaccine or 7 days after killed or subunit vaccine) or (ii) biological mechanism (up to 8 weeks for immune-mediated events). Even though in some cases, and based on their personal experience, clinicians may think that some adverse events are most prob-

ably caused by the intervention, and even take action to guard the patient's health based on this assessment, they nonetheless must report any event occurring in the respective corresponding time frame. In that way, records accumulated from many clinicians may be reviewed by safety committees, where evidence towards causality establishment will be reviewed and policy recommendations, based on unbiased evidence, can be made.

Reports of AEFIs are important elements in the assessment of safety of vaccines and play a major role in public health policy. For vaccination campaigns to be effective the general population needs to be adequately informed so that they maintain confidence in and trust individuals responsible for managing vaccination efficiency and safety [25]. As shown on Figure 2.2, prior to market approval, vaccines are rigorously tested for efficacy and safety, through randomized clinical trials. However, the focus of those trials is efficacy, particularly in the case of widespread, easily transmissible, infections such as influenza where it is hard to fully assess safety due to the limited number of subjects. Additionally, these trials introduce multiple biases:

- They concentrate on a specific subset of the population, and often do not account for variability in gender, age, race, etc. as per their inclusion/exclusion criteria.
- They cannot detect rare adverse events, the cohort of subject enrolled being restricted in size.
- They are limited in time, and will not be able to detect those events for which there is a longer onset period.

Effects in the larger population and in specific subpopulations such as children, pregnant women and the elderly can only be studied post-licensing. Chronic effects, or effects of concomitant administration of other drugs, become evident only after several years of surveillance. As a consequence, there is a need to encourage long-term, widespread post-licensing surveillance. Generally, spontaneous reporting systems are used to monitor for adverse effects in the general population [26]. Each report includes information about adverse events that are at least temporally linked to the vaccination process. Some of those events are causally associated with the vaccine (e.g., it is known that reactions such as rash at the injection site are caused by injectable vaccines), some may or not be related (e.g., patient experiencing a loss of consciousness 3h post-vaccination) and some are probably coincidental (e.g., worker being injured by metallic shard). Analysis of events in large collections of AEFI reports aims to identify signals highlighting differences in frequency of events after administration of a certain vaccine (e.g., a seasonal influenza vaccine), or in certain populations (e.g., children under the age of 2). When such signals are detected health authorities use that

information to prompt investigation of a risk of potential safety issues. Depending on their findings, health officials can make choices such as withdrawing the vaccine from general use or mandating further clinical studies.

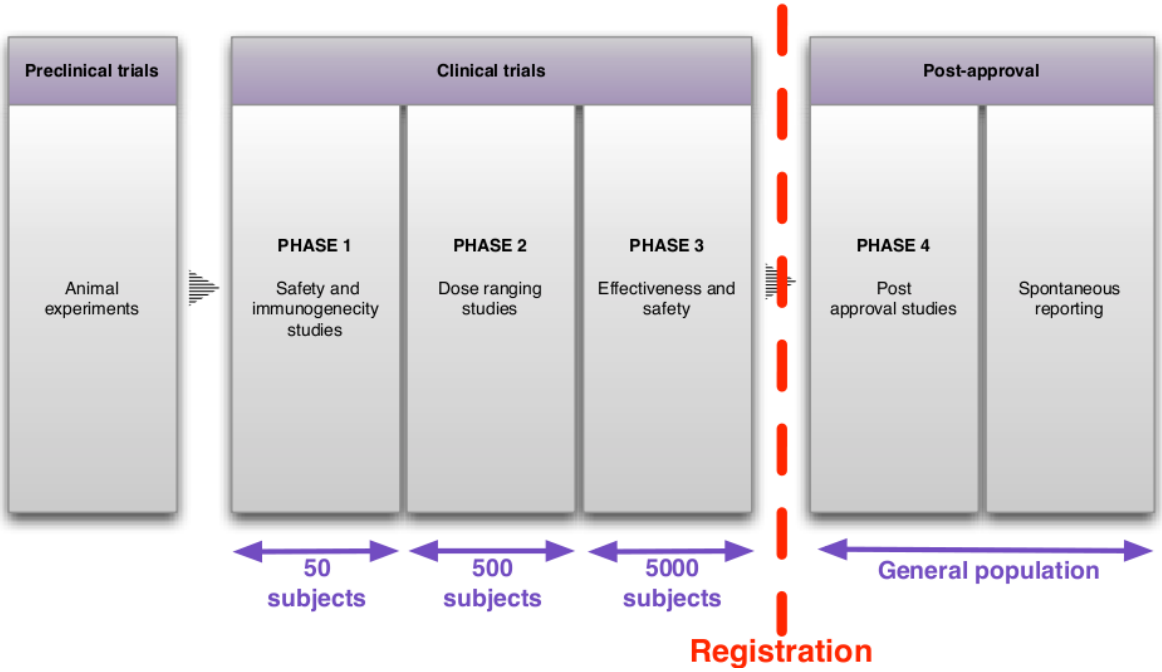


Figure 2.2: The vaccine approval process. During clinical trials, only limited numbers of subjects are studied, warranting need for observation of the vaccine effectiveness and safety in the general population post market approval (numbers are as an example only).

Opposition to vaccination is not new and has existed since the first vaccination against smallpox. Poland and Jacobson [27] detail how it is today more than ever an issue to be contended with, and how efficient reporting and public information, will contribute to defeat anti-vaccination campaigns. One of the most famous roots of the vaccine controversy can be traced to the 1998 Lancet article, "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children" [28], retracted 12 years later due to fraudulent research. This article concluded by demonstrating a causality relation between the Measles, Mumps, and Rubella (MMR) vaccine and autism in children. A survey of 5000 internet users in February 2010 (immediately after the retraction of the Lancet paper) shows that sixty-five percent of Canadian women and seventy two percent of Canadian men surveyed believed a) that the vaccine was unsafe or else b) they were unsure whether or not the MMR vaccine could cause autism. This fear of vaccine side effects has dramatic consequences, causing a drop in vaccination rates in the population. Herd immunity oc-

curs when enough people are vaccinated and provide protection for individuals, such as newborns, who have not yet developed immunity. Partly due to parental refusal of vaccination, herd immunity for vaccine-preventable diseases, such as pertussis, is now compromised [29]. In 2010, 9,143 cases of pertussis (including ten infant deaths) were reported throughout California [3] - the state's worst whooping cough epidemic in 50 years. The importance of pharmacovigilance as a tool for global health policies has been well described [30], even more so in vaccine risk communication, which has been shown [25] to have a direct impact in decisions to immunize in the general public. The resulting drop in vaccination rates is a probable underlying cause in the recent resurgence of vaccine-preventable diseases such as pertussis [29] or even the recent (September 2013) measles outbreaks [31, 32].

Reporting issues - the case for standard guideline representation

Reports are collected from a multitude of sources - physicians and nurses from many different practices, coming from many training backgrounds. Variability in report quality [33] and size of the dataset [3] are significant challenges to derive *adverse event signals* - situations which, with some regularity, predictably lead to some kind of adverse effect - from them. In North America, the assessment of reports is performed by medical officers at the national level, who have no access to the patient and need to rely on the information reported from the primary point of care, which is error prone and may lead to information being missed or erroneously interpreted.

This assessment is often done following a clinical guideline, a protocol[4] that has the objective of guiding medical decision making, assessing patient state, and determining diagnosis or giving treatment. Clinical guidelines might be deployed in at least two places: when first assessing the patient, guiding what should be reported about the patient and how; or when reviewing the reports, guiding how to interpret them together, adjusting for differences in the reports. In current practice, while some efforts are being made to standardize reporting forms, they focus on the kinds of information to include in a report, but not on the terminology to be used when doing so. For example the VAERS report form [34] specifies that a report should include medications the patient is taking, but not detail how a medication should be encoded. Or it might specify that respiratory condition or severity of condition be recorded, but not indicate that a controlled vocabulary be

[3] http://www.cdc.gov/pertussis/outbreaks.html

[4] The OBI defines protocol as "a protocol is a plan specification which has sufficient level of detail and quantitative information to communicate it between domain experts, so that different domain experts will reliably be able to independently reproduce the process."

used.

However in order to look for patterns of symptoms and medication, one needs to be able to count, for example, how often a given symptom occurs - despite it being described in different ways. Such normalization issues may occur even when a specified terminology is used, in cases where the terms are not well documented, or when different terms can be used to describe an equivalent situation. Finally, whereas a clinician might report a cluster of undesirable conditions a patient has experienced, it can be the case that different sets of conditions come from one underlying disease or condition, the primary reason for reporting. In order to have the most power to detect a safety signal, such diseases or conditions need to be recognized and recorded. For example, a goal of pharmacovigilance is to identify all cases of anaphylaxis (an extreme allergic reaction) in a given population, which can manifest via rashes, swelling of the tongue, difficulty breathing etc. It is important that not only those individual manifestations be recorded, but the primary cause of reporting should be identified also when possible.

Assessment choices differ on which cluster of symptoms signifies an underlying condition, or how reliably the information in a report supports an assessment. Figure 2.3 shows elements from four guidelines that assess whether anaphylaxis has taken place, the application of which will result in different clinical assessments, even for the same condition. While the language of reports currently range from controlled vocabulary such as MedDRA to free text, current controlled vocabularies do not sufficiently constrain the meaning of report [15, 35]. To take a step in remedying these problems, a good standard for describing patient conditions is still needed, as well as a consistent and computable manner of describing criteria expressed using that standard.

2.2 Brighton Collaboration

To allow for comparability of data, it is desirable that a global standard for case definitions and guidelines be used for AEFI reporting [35]. The Brighton Collaboration [36] is a global network of experts that aims to provide high quality vaccine safety information. It has done extensive work towards standardizing the assessment and reporting of adverse events following vaccination [40].

2.2.1 Brighton publications

The case definitions provided by the Brighton Collaboration relate symptoms and signs to assessments of whether a particular type of pathological process has occurred, assigning qualitative levels

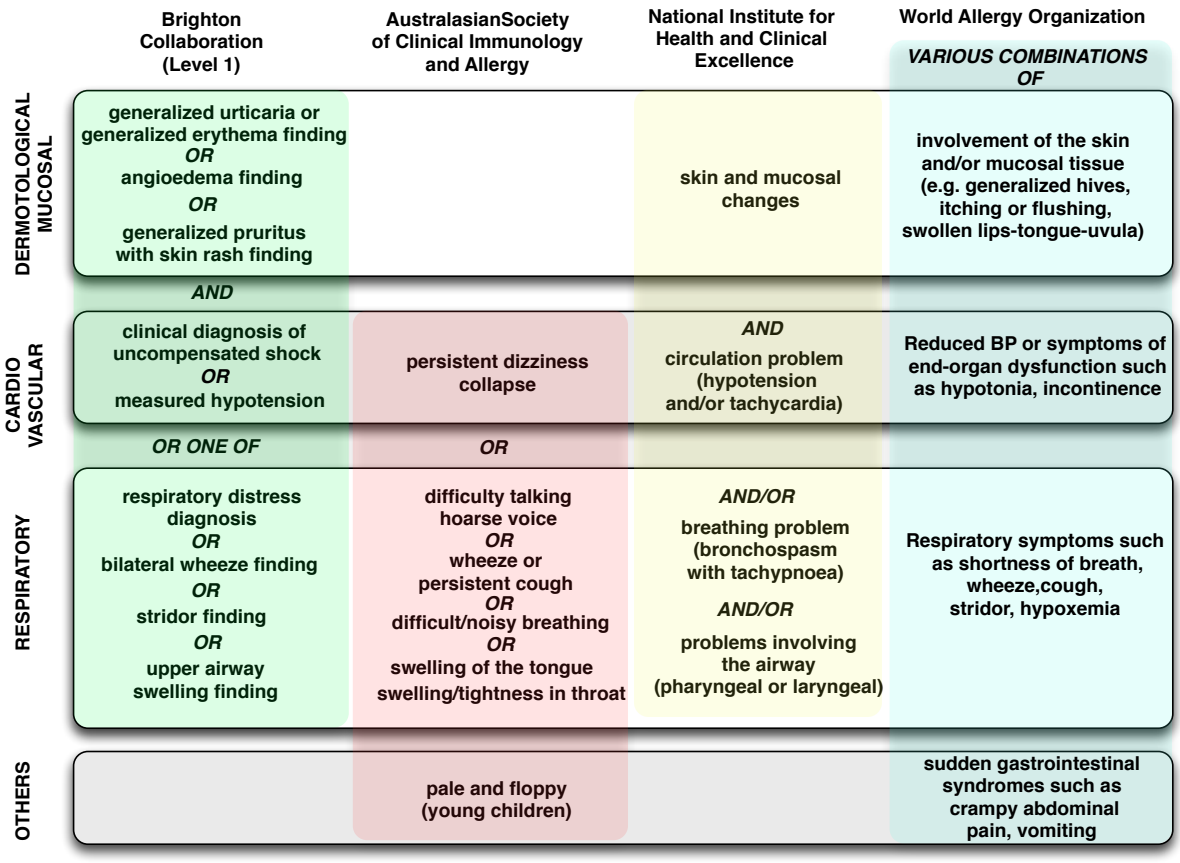


Figure 2.3: Details of some guidelines for anaphylaxis assessment. Horizontal grouping is done by anatomical system in which the manifestation takes place: dermatological/mucosal, cardiovascular, respiratory and others. Colored boxes indicate each of the guidelines considered. Different logical operators (AND, OR, AND/OR) are being used to assemble individual manifestations depending on the guideline considered: Brighton Collaboration [36], Australasian Society of Clinical Immunology and Allergy (ASCIA, [37]), National Institute for Health and Clinical Excellence (NICE, [38]), World Allergy Organization (WAO, [39])

of certainty. They provide guidelines for three activities - data collection, analysis, and presentation of results, aiming to make collected data comparable, informed by the case definitions. By developing and publishing these guidelines, the collaboration creates methodological standards that enable accurate risk assessment. The case definitions neither require, nor assess a causal relation between a given adverse event and the vaccination process. Rather, the case definitions are designed to define levels of diagnosis certainty based on known information about AEFIs.

2.2.2 Automatic Brighton Classification (ABC) tool

The Automatic Brighton Classification (ABC) tool [41] is the only automated classification system that allows users to work with the Brighton case definitions. Given a set of symptoms and a tentative diagnosis, one can confirm the level of diagnostic certainty of an AEFI. Or, given a set of symptoms, the tool can compare them to all Brighton case definitions and report putative diagnoses and their probabilities. Four limitations warrant development of an ontology that would replace the ABC tool:

1. The different signs and symptoms are not defined within the Brighton tool, making it hard at the time of diagnosis confirmation to know if individual findings are those mentioned in the case definitions [15]. While the Brighton guidelines do not provide those definitions, the ontology uses the PHAC glossary ones [42].

2. The tool is embedded within the Brighton portal, and access requires individual login. There are no public API or webservices available, making it not amenable to processing of large amount of data. While there is a mechanism to upload multiple Excel files, this still requires human intervention and raises the issue of sharing medical data with servers located outside of the originating institution.

3. The ABC tool can't be integrated into other systems, and only remote access is available.

4. The rules of classification are hard coded into the ABC tool, which is hard to maintain and extend as new case definitions are being developed [43]. In contrast, ontologies allow for the guidelines to be encoded independently of the application code itself - an update to the ontology does not require updating the business logic of the tools relying on it.

Additionally, use of an ontology allows for text mining of large corpus of data [44], and mapping towards external resources such as MedDRA, which is required when attempting to reconcile existing MedDRA annotations with different guidelines used for their assessment as shown in Chapter 9.

2.2.3 Anaphylaxis according to Brighton

Of special interest for this thesis, the Brighton Collaboration published an anaphylaxis guideline in 2007 [45]. It describes anaphylaxis as "an acute hypersensitivity reaction with multi-organ-system involvement that can present as, or rapidly progress to, a severe life-threatening reaction." The Brighton case definitions have been adopted by the Vaccine Working Group at PHAC and

are captured to some extent in the national Canadian Adverse Event Following Immunization Reporting form [46].

Table 2.2: Case definition of anaphylaxis.

For all levels of diagnostic certainty

Anaphylaxis is a clinical syndrome characterized by

- Sudden onset AND
- Rapid progression of signs and symptoms AND
- Involving multiple (≥2) organ systems, as follows

Level 1 of diagnostic certainty

- ≥1 major dermatological AND
- ≥1 major cardiovascular AND/OR 1 major respiratory criterion

Level 2 of diagnostic certainty

- ≥1 major cardiovascular AND 1 major respiratory criterion OR
- ≥1 major cardiovascular OR respiratory criterion AND
- ≥1 minor criterion involving 1 different system (other than cardiovascular or respiratory systems) OR
- ≥ (1 major dermatologic) AND (1 minor cardiovascular AND/OR minor respiratory criterion)

Level 3 of diagnostic certainty

- ≥1 minor cardiovascular OR respiratory criterion AND
- ≥1 minor criterion from each of ≥2 different systems criterion

Table 2.3: Major and minor criteria used in the case definition of anaphylaxis.

Major criteria	Minor criteria
Dermatologic or mucosal system	
• Generalized urticaria (hives) or generalized erythema • Angioedemaa, localized or generalized • Generalized pruritus with skin rash	• Generalized pruritus without skin rash • Generalized prickle sensation • Localized injection site urticaria
Cardiovascular system	
• Measured hypotension • Clinical diagnosis of uncompensated shock, indicated by the combination of at least 3 of the following: – Tachycardia – Capillary refill time >3 s – Reduced central pulse volume – Decreased level of consciousness or loss of consciousness	• Red and itchy eyes • Reduced peripheral circulation as indicated by the combination of at least 2 of – Tachycardia and – A capillary refill time of >3 s without hypotension – A decreased level of consciousness
Respiratory system	

Major criteria	Minor criteria
• Bilateral wheeze (bronchospasm) • Stridor • Upper airway swelling (lip, tongue, throat, urula, or larynx) • Respiratory distress - 2 or more of the following: – Tachypnoea – Increased use of accessory respiratory muscles (sternocleidomastoid, intercostals etc) – Recession – Cyanosis – Grunting	• Persistent dry cough • Hoarse voice • Difficulty breathing without wheeze or stridor • Sensation of throat closure • Sneezing, rhinorrhea
Gastrointestinal system	
	• Diarrhoea • Abdominal pain • Nausea • Vomiting
Laboratory	
	• Mast cell tryptase elevation > upper normal limit

However, and despite their completeness, the textual, article-like, format of the Brighton case definitions makes it both problematic for clinicians to confirm that they see the relevant symptoms when making the adverse event diagnosis and difficult to automate [15]. As shown in Table 2.2 and 2.3, the anaphylaxis guideline is complex, and this limits its use in pharmacovigilance [15]. Two main barriers for adoption of the anaphylaxis Brighton Collaboration Case Definition (BCCD) have been identified [15]:

1. Health practitioners may not report enough signs and symptoms to allow application of the BCCD,

2. Signs and symptoms terms are not consistently used.

In this thesis, I propose and demonstrate that using an ontology addresses both those issues, by (1) providing logical encoding of the BCCD, which could be used for consistency checking at reporting time, and "prompting" users for missing information (2) providing human readable definitions for terms used in reporting, based on the PHAC glossary [42].

2.3 Ontologies

The project of enabling effective communication and discovery in the biological domain, and pharmacovigilance in particular, is complex. While free-text descriptions can capture relevant experimental details, as exemplified by the methods section of research papers, making the data available for reanalysis and comparison with other related datasets requires a much more systematic and computable approach to capturing information about experiments. In a re-evaluation of 18 peer-reviewed *Nature Genetics* microarray articles, it was reported that the inability of researchers to reproduce analyses was directly linked to data unavailability, incomplete data annotation, or specification of data processing and analysis [47]. It is a significant challenge to unify diverse data sets in a consistent way when the biological relevance of the same entity is labeled differently in different resources, and using a common knowledge representation, such as ontologies, will help provide a stable and consistent context for the information within them [48, 49].

Requirements for a controlled medical vocabulary are described by Cimino in [50], including:

1. **Vocabulary content**: the controlled vocabulary should cover the use cases and domain of knowledge,

2. **Concept orientation**: terms must correspond to at least one meaning ("non vagueness") and no more than one meaning ("non ambiguity"), and that meanings correspond to no more than one term ("non redundancy"),

3. **Concept permanence**: a term may be flagged obsolete or deprecated but once created it is never deleted,

4. **Nonsemantic Concept Identifier**: terms should use numerical, non-semantic identifiers,

5. **Polyhierarchy**: allowing "tree walking" (i.e., browsingthe items of a tree via the connections between parents and children) along different paths depending on the information available and the context of access,

6. **Formal definitions**: include textual definitions as well as the logical ones created by the position in the hierarchy,

7. **Terminologies should support inferencing**: humans and computers should be able to draw conclusions from the information captured [51].

Ontologies are formal representations of knowledge with definitions of concepts, their attributes and relations between them expressed in terms of axioms in some well-defined logic [52]. They specifically address requirements detailed by Cimino in his "desiderata". They model a domain of interest, and provide unique unambiguous definition, both human readable and computer amenable, for each of their term. Biomedical ontologies are sets of terms and relations that represent entities in the scientific world and how they relate to each other. Terms are associated with documentation and definitions, which are, ideally, expressed in formal logic in order to support automated reasoning [53, 54, 55]. Ontologies have dramatically changed how biomedical research is conducted. For example, since the Gene Ontology (GO) was first published in 2000 [53], it has been used and cited in more than 2000 peer-reviewed journal articles [56]. Ontologies have been used in various applications, such as gene expression data analysis [53], literature mining [44], and as the underpinning of a semantic web [57]. There are currently more than 150 biomedical ontologies and 700,000 entities in the National Centre for Biomedical Ontology (NCBO) BioPortal `http://bioportal.bioontology.org/`. With new resources continuously being developed, maximizing ontology sharing and interoperability has become a growing concern [58, 59].

In addition to the development of a biomedical ontology covering the domain of adverse event reporting (AERO) (described in Chapter 8), Chapter 6 specifically addresses items from the desider-

ata [50]. The ID policy provides a standard scheme for numerical identifiers and formalizes a versioning system. It also explores a deprecation policy that ensures terms are never deleted, their identifier therefore being unique and maintained. The MIREOT mechanism (see 6.3) I developed ensures only one term is created for each entity to represent by providing a URI sharing strategy. As terms are related to each other by additional relations to subsumptions in the AERO, one can for example browse adverse event information from the set of all adverse events, or selecting only those involving motor manifestations (i.e., *has_part some motor manifestation*). Finally, reasoners such as Pellet [60] can be used to check consistency of the ontology and infer new facts based on the knowledge captured in the hierarchy.

Use of ontologies falls within three main categories [61]:

1. Knowledge management: Annotation of resources (e.g., the Gene Ontology [62] for gene products, or the Mammalian Phenotype Ontology [63] for phenotypic information) that increase recall and precision when retrieving biomedical information. In [64], subclasses of the originally searched taxon were automatically included in results, such that a search for "mammalian models" would return those pertaining to human, mouse etc.

2. Data integration: Ontologies such as TAMBIS [65] facilitate information exchange and semantic interoperability, data integration. While TAMBIS accesses only five resources (Swiss-Prot, Enzyme, Cath, blast and Prosite), modern SPARQL endpoints allow for querying across multiple datasets [66, 67, 68]. In [57], Ruttenberg et al. describe a query that retrieves gene records and the name of signal transduction related processes that the gene products participate in that are related to pyramidal neurons, by querying across multiple data sources hosted on Neurocommons.

3. Decision support and reasoning: For example in [69], an ontology based on the NCI Thesaurus is used to grade glioma tumors automatically and compare the classification with 11 pathology reports.

In this thesis, building an ontology means addressing three essential roles [70]:

1. An ontology in a given domain is a collection of representations of the important types of things in that domain, with an understanding that instance representations any of these types should be considered proxy for things in the world. The truth of assertions made on the representations is judged by the facts about that which the representations serve as proxies.

2. An ontology is an active computational artifact. The assertions that are made are in a subset of first order logic which can be checked for consistency and from which logical consequences can be computed. I aim to take advantage of this by asserting as many axioms as feasible. As a way of improving quality, these axioms maximize the opportunity for consistency checks to turn up errors. Each case report is classified when the specifics of the case satisfy a guideline for the purpose of diagnosis and screening. For example a query can be executed for all the cases of adverse effects that affect a specified anatomical system such as 'skin rash located at some dermatological-mucosal system'.

3. An ontology facilitates scholarly and technical communication.

    - Working in a large community of ontology developers who split the labor and use each other's work, the OBO Foundry [55], and who together work out principles that encourage quality through careful analysis,
    - Mediating communication between clinicians and technical specialists when the practice of having literate documentation about the types in the ontology is followed, so that clinicians and co-workers can have reasonably expectations of what the data means,
    - Being part of the package distributed so that other researchers can reproduce results.

2.4 The OBO Foundry

Biomedical investigations use empirical approaches to investigate causal relationships among a large range of variables. The wide range of possible investigations presents a number of challenges when building tools to describe experimental processes. There are varying levels of complexity and granularity and a wide range of material and equipment is used. Furthermore, the use of varying terminology by different communities makes data integration problematic when representing and integrating biomedical investigations across different fields of study.

The use of ontologies has been successful in biological data integration and representation [71, 72] and there have been multiple efforts to develop ontologies aimed at providing clearer semantics for data (GO [53], FuGO [73], MGED [74], EXPO [75], LABORS [76], MSI ontology [77]). Work in the transcriptomics, proteomics and metabolomics communities has proceeded in parallel, producing ontologies with overlapping scopes. Though each focuses on particular types of experimental processes, many terms, such as investigation and assay, are common to all. Merging common aspects

of these formalisms is useful as it provides a mechanism by which terms can be used and understood by all, reducing ambiguity and difficulties associated with post-hoc attempts to integrate data.

The practice of consolidating representations is endorsed by organizations such as the OBO Foundry [55] which requires all member ontologies to define a term only once among them (orthogonality). OBO Foundry members use a common set of relations from the Relations Ontology (RO) [78] and the upper level Basic Formal Ontology (BFO) [79] in order to facilitate cross ontology consistency and to support automated reasoning [55]. OBO ontologies also adhere to common naming conventions in order to make it easier to learn and understand them: this common metadata set is described in section 6.2.2.

The development of a new biomedical ontology covering a specific domain is often an ambitious, time-consuming project, usually requiring extensive cross-community collaboration [80, 81]. The Open Biomedical Ontologies (OBO) Foundry is an open community that has established a set of principles for ontology development with the goal of creating a suite of interoperable reference ontologies in the biomedical domain [55]. These principles require that member ontologies be open, orthogonal, expressed in a common shared syntax, and designed to possess a common space of identifiers. One way of meeting the goal of interoperability is to reuse existing resources by importing them into the to-be-created ontology. For example, the VO [82], described in Chapter 4, relies on many terms (e.g., administering substance in vivo) already described by other biomedical ontologies, such as the OBI [83]. Authors of resources submitted to the OBO Foundry library[5] commit to working together to increase quality of resources.

As a result of this collaborative work, resources that are part of the OBO Foundry are orthogonal in scope (i.e., each resource describes a specific, non-overlapping domain) - and common policies are devised and followed [84]. To increase interoperability, ontologies use a common upper-ontology (Basic Formal Ontology (BFO))[85] and a common set of relations (RO)[78]. Policy adoption at the level of the OBO Foundry is done by decision of the OBO coordinators, a set of individuals helping build a community adhering to the OBO principles shown in Appendix C. Sharing development principles and domains aims at decreasing workload for ontologies developers, and ensure each domain is covered adequately by experts in the area. The idea of working collaboratively in a "Foundry" type of framework has also been adopted by the MInimum reporting guidelines for Biological and Biomedical Investigations (MIBBI) project [86], with goals similar to the OBO Foundry.

[5] `http://www.obofoundry.org/`

In addition to creating a suite of reference ontologies, the OBO Foundry also promote their use in the annotation of multiple datasets in the interest of enabling effective integration of data in this field. Biomedical ontologies are, typically, consensus-based controlled biomedical vocabularies of terms for classes and relations associated with natural language definitions and logical axioms formulated to promote automated reasoning. A key challenge in establishing such consensus and reaping the consequent benefits of widespread use is the wide dissemination of the terms from these ontologies making them discoverable, understandable, and (re)usable. Chapter 7 describes the Ontobee server that was implemented by my collaborators at the University of Michigan in the context of the OBO Foundry for this purpose.

In the context of the OBO Foundry, resources are developed in a modular way, as shown in Figure 2.4. A top-level ontology, BFO, provides the basic scaffolding on which others can build. BFO describes high-level entities, such as occurrents (those things that occur at some time, such as *processes*), continuants (those things that perdure through time, such as *material entities* or their attributes, such as *qualities*). Other resources built under BFO with the goal of providing representations for specific domains. For example, OBI extends bfo:processes encompass assays, and material or data transformation in the domain of biomedical investigations. Organisms are subclasses of material entities, and bear different roles, such as *principal investigator* or *specimen*, or qualities, such as *radioactive*. Similarly, OGMS and VO, described in Chapter 4, aim at representing clinical and vaccine data. Finally, AERO is discussed in Chapter8 for application to adverse events.

2.5 OWL and the Semantic Web

2.5.1 The Semantic Web

Integrating heterogeneous data from multiple sources or databases is a well-known problem [87, 88]. With the advent of the Web, there is an additional need to unify multiple data sources and serve the end user with a unified view they can browse and query [89, 90]. The semantic web aims at extending the existing web of documents into a web of data designed to also be processed automatically. It relies on providing unambiguous names for things, such as classes and relationships between them, that are well organized and documented in ontologies. Data is expressed using standard knowledge representation languages, such as Resource Description Framework (RDF) [91], OWL [92], and Rule Interchange Format (RIF) [93] and can be queried using for example the SPARQL Protocol and RDF Query Language (SPARQL) [94]. This enables computationally assisted exploitation

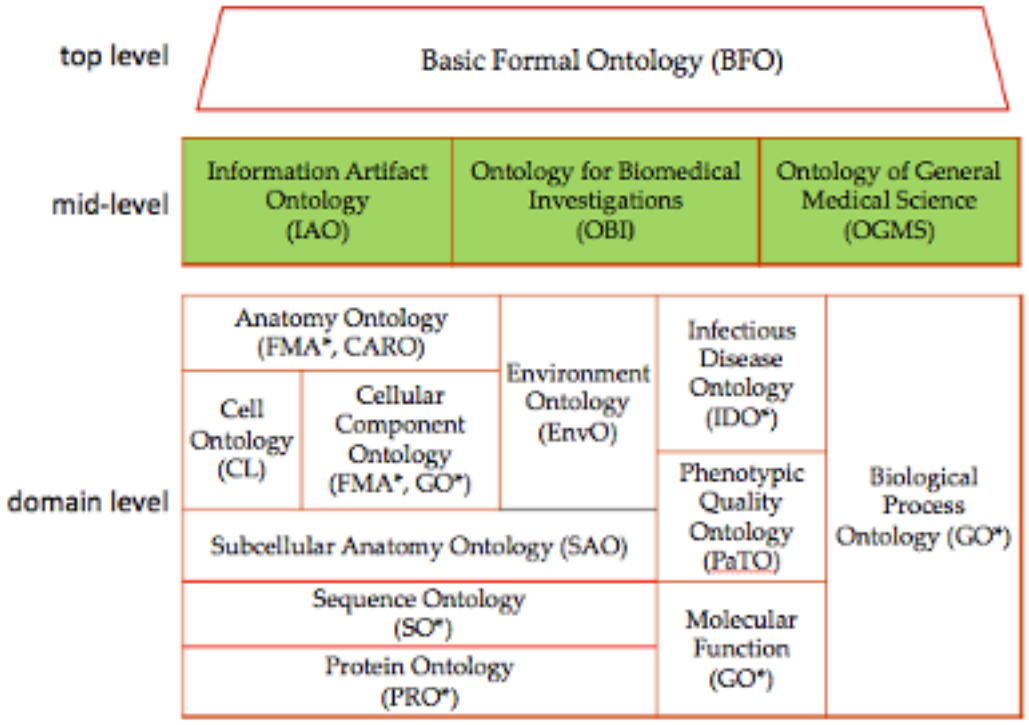


Figure 2.4: Coverage and distribution of OBO Foundry ontologies. Figure reproduced from `http://ontology.buffalo.edu/obofoundry/Graz2012/1-The_OBO_Foundry.ppt` by Barry Smith, licensed under CC-by-nc-sa.

of information: machines can work with data, allowing for consistency checking, querying and inferences over the datasets. Finally, data is integrated from different sources. As described in [95], "The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data." A variety of systems are imagined to benefit from a web of data, ranging from agents that understand enough of such data to reliably act on behalf of their users, as well as systems that are able to discover previously unknown relations among data in such a web of data [95].

In his linked data note [96], Sir Tim Berners-Lee highlighted four underlying principles (with minor paraphrasing):

1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up (i.e., dereference) those names.

3. When URIs are dereferenced, provide useful information, using the standards (RDF, SPARQL).

4. Include links to other URIs in that useful information, so that more things can be discovered.

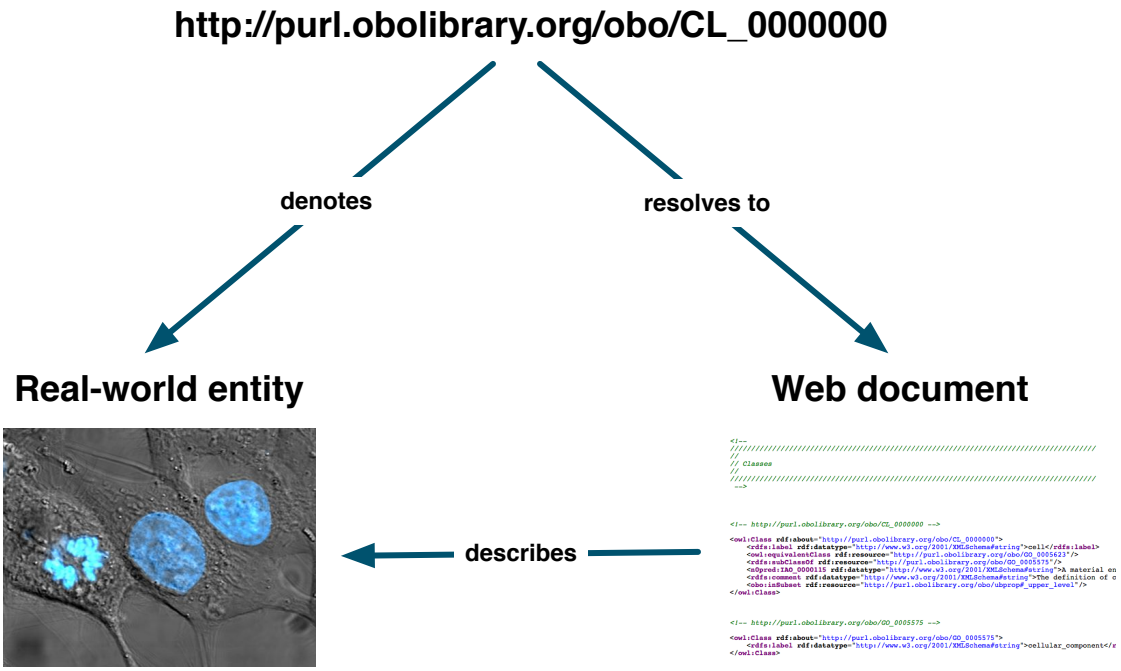


Figure 2.5: Meaning of URIs. The URI `http://purl.obolibrary.org/obo/CL_0000000` denotes cells in the real world. This URI can be dereferenced (e.g., via a web browser) and upon doing so provides useful information using RDF. This information may contain links to other relevant resources.

This "meaning of URIs" is depicted in Figure 2.5. The Linked Open Data (LOD) [68] data cloud includes Bio2RDF [66], Uniprot [97], DBPedia [98], and Neurocommons [57], and aims at integrating even more of those datasets. Bio2RDF encompasses information from many key bioinformatics resources (e.g., KEGG [99], PubMed [100], Uniprot [97]). This allows for example the querying of the Uniprot dataset using a PubMed node identifier, because the identifier of the PubMed resource is shared between the two datasets. Figure 2.6 illustrates the architecture of the semantic web. Reusing the technologies it relies upon provides some of the building blocks for data sharing in the biological domain. Logical languages such as RDF and OWL have effective tool implementations, such as HermiT [101], Pellet [60] or Fact++ [102]. Additionally, there is no need to design individual "data models" - a common source on non-integrable data - as a standard one is provided by RDF, as well as databases supporting it, such as Virtuoso [103], Stardog [104], OWLIM [105], or

Sesame [106]. Finally, SPARQL [94] is a standard query language for RDF.

Section 6.2.1 describes how I implemented the linked data principles by formalizing a common ID policy normalizing use and format of HTTP URIs across the OBO Foundry, while Chapter 7 details the dereferencing mechanism implemented as default for resources relying on this ID scheme. Throughout this thesis, resources have been built in OWL, to which the next section provides an introduction.

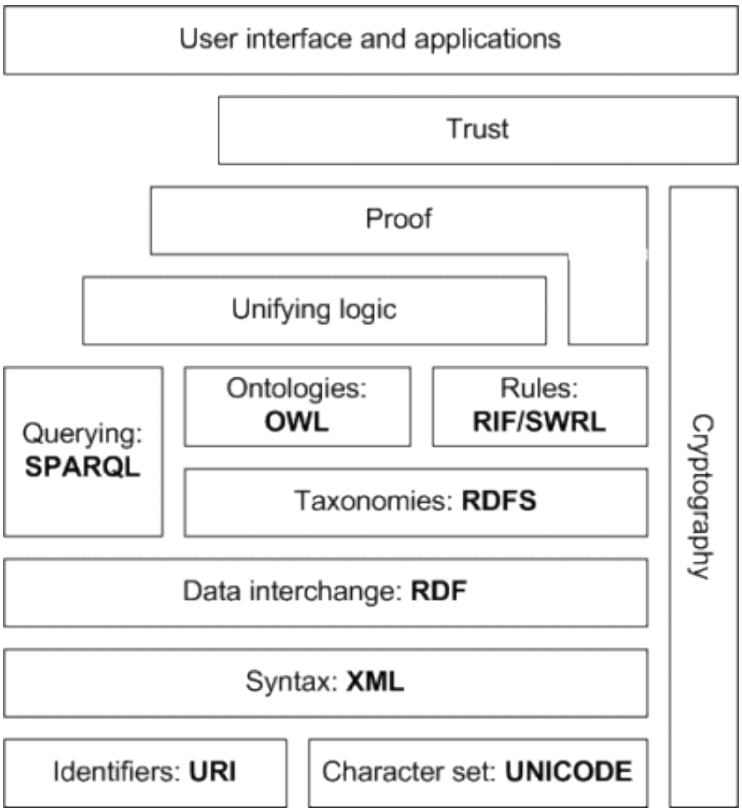


Figure 2.6: The Semantic Web architecture, as described by Tim Berners-Lee, `http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html`. Image in the public domain, from wikipedia.

2.5.2 Components of an ontology

Ontology languages

As shown on Figure 2.6, RDF, RDF Schema (RDFS) and OWL are core components of the semantic web. In RDF, data is represented by triples - a set of subject, predicate, object, in which each entity can be denoted by its corresponding Uniform Resource Identifier (URI). Figure 2.7 details how the triple *eukaryotic cell has_part nucleus* can be encoded in RDF.

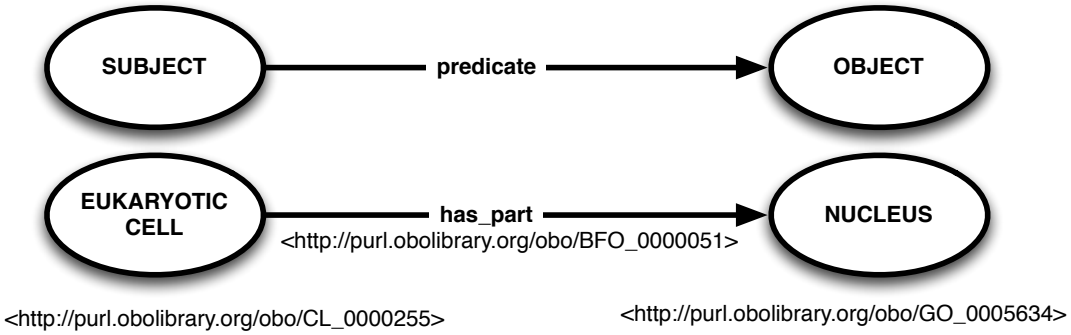


Figure 2.7: Encoding of triples in RDF

RDFS allows the addition of simple relations to RDF, such as *rdfs:subclassOf*, which supports basic inference. For example, by asserting the following triples subject, predicate, object: *animal cell rdfs:subclassOf eukaryotic cell* and *eukaryotic cell rdfs:subclassOf cell*, a reasoner can infer that *animal cell rdfs:subclassOf cell*. OWL provides the higher level of expressivity. For example, axioms such as *eukaryotic cell disjointWith prokaryotic cell* cannot be expressed using RDFS [107]. However, the more expressive a language is, the more difficult it is to reason over representations built using that language [108]. OWL was chosen as a means to provide a balance between expressivity required, computation capability and tool support. The OWL DL subset was indeed suitable to represent the axioms required (described in Chapter 8), while being computationally decidable [109]: reasoners are guaranteed to finish and produce a result, though in practice the performance is not guaranteed, i.e., the memory space and time required for computation can be infinite. In Chapter 9 I discuss some reasoning issues encountered during this thesis, and solutions I implemented to address them.

While many formats can be used to encode OWL files, the Manchester OWL Syntax [110] was chosen as a more user friendly option for examples in this dissertation.

Ontology structure

An ontology can be split into a TBox and ABox [111]. The TBox, or **T**erminological box, contains intensional knowledge: classes and relationships amongst them, which represents the general knowledge about a domain. For example, a woman can be defined as a female person. On the other hand, the ABox, or **A**ssertions box, encodes the extensional knowledge, which is specific to a domain, via instances. For example, the individual ANNA is a female person [112]. While the TBox is often compared to relational databases schemas, the ABox can only approximately be seen as the set

of instances in the database. Relational databases store a finite amount of data, and data that is not present is simply inferred to negative. However in OWL, the Open World Assumption (OWA) prevails, and data that is not explicitly stated as negative is rather considered unknown [113]. For example, if the statement ANNA is a female person is made, and the question"is Mary a female person?" is asked, a closed world (such as SQL) system will answer "No", while an open-world system will answer "unknown". As a consequence, it may be required in some cases to add negative axioms to represent knowledge extracted from a database, as shown in Chapter 9. Additionally, ontologies contain documentation, in the form of metadata either on the classes, properties and instances in the model, or on the model itself. Details of the shared metadata set I developed for the OBO Foundry is in Section 6.2.2 for annotation on classes, properties and relations, while Section 6.2.1 describes some ontology metadata such as versioning information. Finally, ontologies can reuse each other via the owl:imports statement, which allows to reference another OWL ontology containing definitions, whose meaning is considered to be part of the meaning of the importing ontology [114]. A discussion of the import mechanism and limitations is available in Section 6.3.

Classes

In OWL, classes are set of individuals called the class extension [114]. Classes can be described by either a class name (or URI) or a description of its extension (e.g., all cells that have a nucleus are eukaryotic cells, or cells are prokaryotic or eukaryotic). In OWL, 6 descriptions are allowed [114]:

1. A class identifier
2. An enumeration of individuals (using owl:oneOf)
3. A property restriction (value or cardinality constraint)
4. An intersection of classes descriptions (AND)
5. An union of classes descriptions (OR)
6. A complement of a class description (NOT)

Classes descriptions can be turned into classes axioms using one of the three constructors:

1. rdfs:subclassOf: the subject class extension is a subset of the parent class extension. For example, animal cell subclassOf cell.

2. owl:equivalentClass: the subject and object class extensions are identical. For example 'eukaryotic cell' and 'cell that has a nucleus'.

3. owl:disjointWith: the subject and object class extensions share no common member. For example 'eukaryotic cell' and 'prokaryotic cell'.

Properties

OWL defines different types of properties [114]. Object properties link instances to instances, while data properties link instances to data values. OWL DL further specifies annotation properties on class, individuals, properties provided some conditions are respected (for example, annotation, object and data properties must be disjoint). In the context of the OBO Foundry, common properties used to be included within the RO [78]. Common relations have been now included in the upcoming version of BFO, BFO2.0.

Individuals

Individuals are the members of a class extension. A reasoner can infer their class membership via asserted facts on those individuals. For example, if one asserts that in their current experiment the cells C1 have a nucleus, then a reasoner could infer that C1 is a member of the class extension of "eukaryotic cell". Such facts, allowing classification of individuals into their respective types, are necessary and sufficient conditions (it is necessary AND sufficient to have a nucleus to be an eukaryotic cell[6]). Other types of facts, necessary conditions, are required but don't allow for type inference. For example, all bone cells are part of the bone element, but not everything that is part of the bone element is a bone cell. Finally, when individuals are asserted or can be inferred as members of two disjoint classes, an inconsistency occurs [115].

[6]As an interesting case, red blood cells are eukaryotic cells despite losing their nucleus during maturation. A temporal qualification of existing relations, such as *has part* is being worked upon in the context of the development of BFO2.

Chapter 3

Representing biomedical investigations

3.1 Introduction

Before being able to describe pharmacovigilance processes accurately and in a standard way, a general framework for biomedical investigations is required. In this chapter, I describe parts of the OBI development in Section 3.2, and how it can be extended to represent vaccine protection assay and statistical analysis thereof. I was a coordinator for the OBI consortium and the core developer in charge of most development at this time, and participated in implementation of the models: all made use of the infrastructure I developed. Coordinators make decision on general guidance of the OBI, while core developers are directly concerned with the editing and implementation required. I produced the release OWL file on which this chapter is based. My work focused on the neuroscience investigation (Use case 1) in collaboration with Dirk Derom and Alan Ruttenberg, as well as the vaccine protection investigation (Use case 2) in collaboration with Yongqun He. OBI is applied to model, among others, an investigation of vaccine protection against influenza viral infection. The vaccine protection investigation measures how efficient a vaccine or vaccine candidate is at inducing protection against a virulent pathogen infection in vivo. Section 3.3 then describes the application of the vaccine protection pattern to the ANOVA analysis of variables involved in a *Brucella* vaccine protection efficacy.

3.2 The Ontology for Biomedical Investigations (OBI)

The OBI Consortium[7] is developing an integrated ontology for the description of biological and clinical investigations. This includes a set of 'universal' terms that are applicable across various biological and technological domains, as well as domain-specific terms. OBI supports the con-

[7] http://obi-ontology.org/page/Consortium

sistent annotation of biomedical investigations, regardless of the particular field of study. The ontology represents the design of an investigation; the protocols, instrumentation and material used; the data generated; and the type of analysis performed. OBI also represents roles and functions used in biomedical investigations. OBI has been used in experimental investigations in different communities, for example, Bioinvindex (`http://www.ebi.ac.uk/bioinvindex`), isa-tools (`http://isatab.sourceforge.net/`), and IEDB (`http://www.immuneepitope.org/`).

OBI defines an investigation as a process with several parts, including planning an overall study design, executing the designed study, and documenting the results. An investigation typically includes interpreting data to draw conclusions. Biomedical experimental processes involve numerous sub-processes, involving experimental materials such as whole organisms, organ sections and cell cultures. These experimental materials are represented as subclasses of the BFO class material entity. OBI uses BFO's material entity as the basis for defining physical things:

- A material entity is an independent continuant, a continuant that is a bearer of quality and realizable entity(s), in which other entities inhere and which itself cannot inhere in anything [79].

- Material entities are entities that are spatially extended, whose identity is independent of that of other entities, and which persist through time, for example organism, test tube, and centrifuge.

- Material entities can bear roles, typically socially defined, which are realized in the context of a process, e.g., study subject role, host role, specimen role, patient role; and functions, results of design or evolution that depend on their physical structure e.g., measure function, separation function and environment control function. The function is considered to inhere in the material entity and be realized by the role that material entity plays in a process.

To assess the completeness of the OBI release and demonstrate the use of OBI for annotation, two representative use cases are presented. These demonstrate how to model entities and relations between entities involved in experimental processes using OBI. The first use case models a neuroscience experiment described in a journal article [116] and shows how logical definitions are constructed using parts of external ontologies imported into OBI. The second use case details how OBI is used to model vaccine studies. Having the ability of integrating across multiple domains is of particular interest for this thesis: as an example, consider that a vaccine candidate against

Alzheimer disease may induce specific changes on the brains of transgenic mice or human patients [117]. Therefore enabling queries across the domains of vaccinology and neuroscience would be of utility in conducting such research.

3.2.1 Use case 1: Neuroscience investigation

This investigation studied the role of the primate caudate nucleus in the expectation of reward following action [116]. While the caudate nucleus responds preferentially to eye movements in different directions, the response begins prior to eye movement and is dramatically increased when there is expectation of reward for the preferred direction. Here a single trial is represented, in which the visual target, a light, is presented to the animal and the neural response is recorded as data. This single trial model contains two processes (Figure 3.1):

1. Stimulating monkey with a light source, which is an example of presentation of stimulus. The Japanese macaque monkey participates as the subject and light source as the stimulus, during the process of a measuring neural activity in the caudate nucleus assay.

2. Measuring neural activity in the caudate nucleus: this process is a subclass of the process extracellular electrophysiology recording, which unfolds in the caudate nucleus that is part of the Macaca fuscata, of which the Japanese macaque monkey is an example. The anatomical term caudate nucleus is imported from the Neuroscience Information Framework standardized (NIFSTD) ontology [118] and used in the logical definition of the assay.

The light on the tangent screen here is a light source used to present the stimulus to the study subject. The function of the microelectrode, part of the single unit recorder (an example of processed material), is realized in the measuring neural activity in the caudate nucleus process. The process measuring neural activity in the caudate nucleus has the specified input a neuron and the specified output a neuronal spike train datum.

3.2.2 Use case 2: Vaccine protection investigation

A vaccine protection investigation (also known as a vaccine challenge experiment) measures how efficiently a vaccine or vaccine candidate induces protection against a virulent pathogen infection in vivo. Figure 3.2 demonstrates how to use OBI to represent a typical vaccine protection investigation via the following three sub-processes:

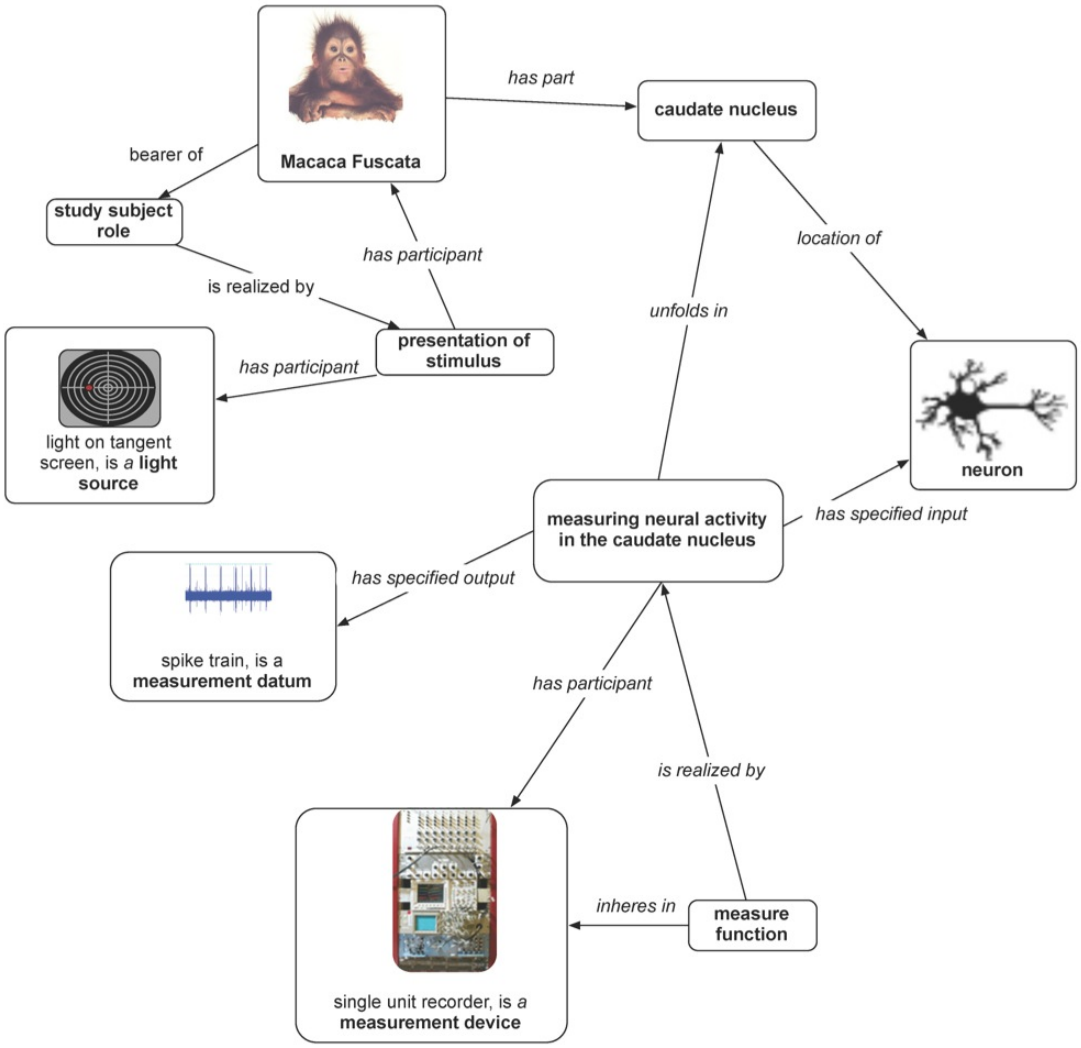


Figure 3.1: OBI modeling of a single trial in the neuroscience study (a fragment). In this and subsequent figures, boxes represent instances, labeled by the class they are instance of and relationships as links labeled in italics. In several cases the parent class is also noted with the class label. Note that in typical use only some instances would be explicitly created - others would be inferred as a consequent of OBI's definitions. Some processes in this experimental trial are presentation of stimulus, measuring neural activity in the caudate nucleus, and stimulating monkey with light source. Some continuants are *Macaca fuscata*, study subject role, spike train

1. A vaccination is a kind of administering substance in vivo process that realizes some material to be added role, borne by a vaccine (e.g., VacX) as well as a target of material role borne by an organism that also bears a host role (e.g., mouse). The term vaccination is a term

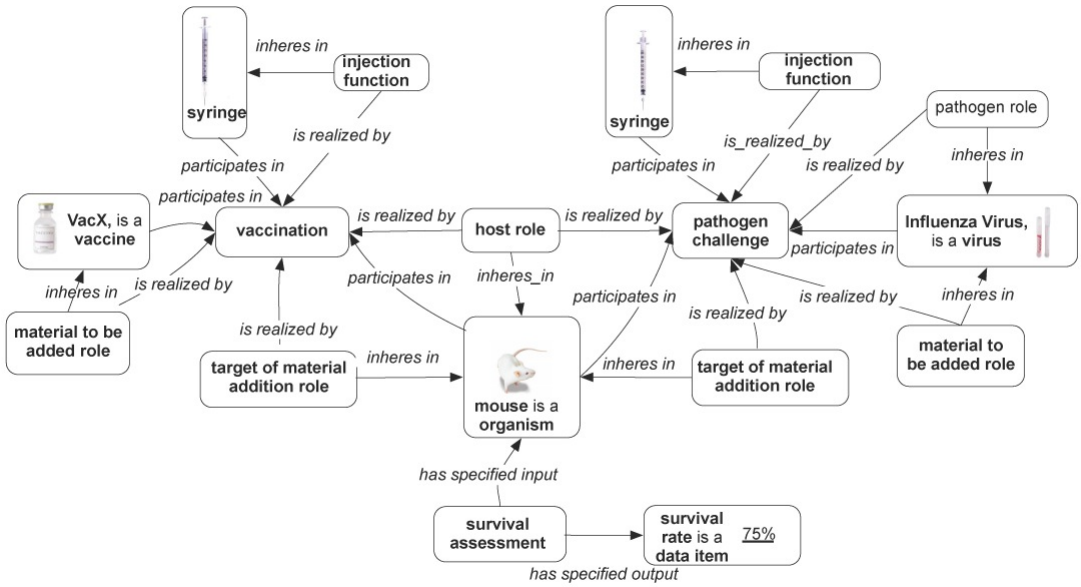


Figure 3.2: OBI modeling of vaccine protection investigation (a fragment). Major processes are vaccination and pathogen challenge, both of which are subtypes of administering substance in vivo. The roles target of material addition and material to be added role are defined with respect to this parent class. Some objects are syringe, mouse, host role, target of material addition role, VacX and a portion of Influenza Virus. Note that while the figure shows a single input for the survival assessment, in fact there would be many replicates of the experiment shown, with observations of mouse survival from all of them input to the survival assessment.

imported from the Vaccine Ontology [119]. The vaccination process realizes the injection function inhering in a syringe (itself a processed material).

2. A pathogen challenge is also a kind of administering substance in vivo process. It realizes a number of roles - a pathogen role and material to be added role borne by the challenge organism (e.g., Influenza Virus), and a target of material role and host role borne by another organism (e.g., mouse). An injection function that inheres in a syringe is realized by the pathogen challenge process.

3. A survival assessment is an assay that measures the survival rate (occurrence of death events) in one or more organisms that are monitored over time. The survival assessment is a protection efficiency assay that has specified input a number of organisms (e.g., mouse) and has specified output a survival rate, in this case a measurement datum that records that 75% of mice

survived the pathogen challenge.

3.2.3 Discussion

OBI was built to provide a comprehensive and versatile representation of biomedical investigations. In the three biological use cases above, individual experimental steps - the two processes in the neuroscience use case, the three processes in the vaccine protection case, and the three processes in the functional genomics case - all fall under planned process in OBI.

In the example of the neuroscience investigation use case, the construction of logical definitions of the experimental process prompted questions to domain experts, because details to capture were not explicit in the publication. For example, was the location of the micro-electrode extra- or intra- cellular? Was all spike train data recorded from the caudate nucleus? How does a spike train relate to the GO biological process regulation of action potential [GO:0001508]? Based on the answers, OBI's existing assays were augmented, and several terms from external ontologies, for example NIFSTD, were imported. When relations not yet present in OBI were needed, rather than define them *de novo* relations from ro_proposed ([`http://obofoundry.org/cgi-bin/detail.cgi?ro_proposed`](http://obofoundry.org/cgi-bin/detail.cgi?ro_proposed)) were used. For example, unfolds in specifies that an occurrent (process) happens in a certain location (i.e., the assay of spike trains in the caudate nucleus). Finally, the NCBI taxonomy [120] was used to describe the species involved in this experiment. As described further in Section 6.3, re-use of external resources fulfills two purposes. First, as domain experts have already devoted time to defining terms in these external ontologies substantial efforts are prevented by not replicating that work. Second, by re-using existing resources that others already use, the potential for future data integration is improved, by making it unnecessary to map between different identifiers denoting the same entity.

In developing the neuroscience use case some decisions about choosing an appropriate level of detail were challenging: in this use case instances of the classes were not included. Instead focus was on adding classes that can be re-used for other use cases and communities. The analysis and the classes defined can then serve as design patterns for other neuroscience assays. Depending on the use case, OBI intends to be able to model the desired level of details (granularity), from molecular level experiments to higher level of biomedical investigations. OBI can be used at a more or less granular level depending on the user community needs.

In the second use case, the vaccine protection investigation includes three processes. The processes vaccination and pathogen challenge are disjoint subclasses of administering substance in

vivo. The process survival assessment is a type of assay (Table 3.1). All these required processes, as well as all other entities described in the use case could be represented using OBI idioms. Syringe is a processed material that participates in different processes. Entities such as vaccine are types of material entity. Host role, pathogen role, and material to be added role are types of roles.

Table 3.1: Ontology terms used in the use cases (note: instances are not included).

Ontology terms	**Sources and term IDs**	**Parent class**	**Use cases**
Classes			
administering substance in vivo	OBI: OBI_0600007	material combination	2
assay	OBI: OBI_0000070	planned process	1
caudate nucleus	NeuroLex: birnlex_1373	anatomical entity	1
extracellular electrophysiology recording	OBI: OBI_0000454	assay	1
function	snap#Function	realizable entity	1,2
host role	OBI: OBI_0000725	role	2
IndependentContinuant	continuant		
injection function	OBI:OBI_0005246	function	2
light source	OBI: OBI_0400065	processed material	1
Macaca fuscata	NCBI_Taxon: NCBITaxon_9542	organism	1
material combination	OBI: OBI_0000652	planned process	2
material to be added role	OBI: OBI_0000319	role	2
material entity	snap#MaterialEntity	Independent continuant	1,2
measure function	OBI: OBI_0000453	function	1
measurement device	OBI: OBI_0000832	processed material	1
measurement datum	IAO: IAO_0000109	data item	1,2

Ontology terms	**Sources and term IDs**	**Parent class**	**Use cases**
measuring neural activity in the caudate nucleus	OBI: OBI_0000812	extracellular electrophysiology recording	1
micro electrode	OBI: OBI_0000816	processed material	1
neuron	FMA: FMA:54527	anatomical entity	1
organism	OBI: OBI_0100026	material_entity	2
pathogen challenge	OBI: OBI_0000712	administering substance in vivo	2
pathogen role	OBI: OBI_0000718	role	2
presentation of stimulus	OBI: OBI_0000807	process	1
process	span#Process	processual entity	1,2
processed material	OBI: OBI_0000047	material entity	1,2
role	snap#Role	realizable entity	1,2
spike train datum	OBI: OBI_0000801	measurement datum	1
study subject role	OBI: OBI_0000097	role	1
survival assessment	OBI: OBI_0000699	assay	2
survival rate	OBI: OBI_0000789	measurement datum	2
syringe	OBI: OBI_0000422	processed material	2
target of material addition role	OBI: OBI_0000444	role	2
vaccination	VO: VO_0000002	administering substance in vivo	2
vaccine	VO: VO_0000001	material entity	2
Property terms			
bearer_of	RO: OBO_REL#bearer_of		1
has_participant	ro.owl#has_participant		1

Ontology terms	Sources and term IDs	Parent class	Use cases
has_specified_input	OBI: OBI_0000293		1,2
has_specified_output	OBI: OBI_0000299		1,2
inheres_in	RO: OBO_REL#inheres_in		1,2
is_a	RO: OBO_REL:is_a		1,2
is_realized_by	IAO: IAO_0000122		1, 2
location_of	ro.owl#location_of		1
part_of	ro.owl#part_of		1
unfolds_in	RO: OBO_REL#unfolds_in		1

That OBI can be used to represent experimental processes for different applications and domains is appealing because it suggests that biomedical investigation work can be better leveraged. For the domain of vaccine investigation, approximately 400 vaccines have been manually curated and stored in the Vaccine Investigation and Online Information Network (VIOLIN; `http://www.violinet.org`) vaccine database system [121], described in Chapter 4. Currently, the vaccine protection experimental data in VIOLIN is stored in plain text and can be difficult to interpret. The lack of a common ontology to aid in representing this data has prevented optimal use of the VIOLIN vaccine data. Applying the representation described above to that data would enable advanced querying both within the data as well as across data from other biomedical communities that represent their data using OBI.

3.3 Ontology representation and ANOVA analysis of Brucella vaccine protection investigation

Brucella is an intracellular bacterium that causes brucellosis, the most common zoonotic disease worldwide. Vaccine challenge studies are only performed in animal models, and typically occur at the preclinical stage. They are critical in determining whether a vaccine can yield the desired immune response. In this section, it was hypothesized that some experimental variables significantly contribute to *Brucella* vaccine protection efficacy while others do not. To investigate this hypothesis, the vaccine protection investigation was represented using VO and OBI. This model was then evaluated by my collaborators at the University of Michigan using literature-curated data.

3.3.1 Methods

The following methods were applied in this study:

1. **Ontology representation of ANOVA Statistical analysis:** The analysis of variance (ANOVA) was modeled primarily in OBI. A design pattern was generated. The use case in this study is ANOVA in terms of a linear model.

2. **Ontology-based representation of vaccine protection investigation:** All variables in this use case are represented using different ontologies as needed. The main ontologies used include VO, OBI, and IAO.

3. **Literature curation of individual *Brucella* vaccine protection data:** Peer-reviewed *Brucella* vaccine protection research papers were obtained from PubMed search. These papers were manually curated to identify variables and extract values taken by these variables potentially important for vaccine protection efficacy investigation. The data were stored in an OWL file.

4. **Ontology-based ANOVA analysis of *Brucella* vaccine protection results:** ANOVA was applied to study the *Brucella* vaccine protection investigation instance data. The results were also represented in the ontology.

I performed the ontology implementation (items 1 and 2), while my collaborators at University of Michigan executed items 3 and 4.

3.3.2 Results

Ontology design pattern of ANOVA data analysis

The analysis of variance (ANOVA) provides a statistical test of whether or not the means of several groups are all equal. In statistics, ANOVA includes a collection of statistical models (e.g., linear models), and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. The ontology-based ANOVA data analysis design pattern is illustrated in Figure 3.3. ANOVA is a subclass of data transformation process in OBI. F-test is part of ANOVA process. ANOVA has specified input some data item, which come from two sources. They can be the output of individual processes (e.g., CFU reduction assay) or of a discretization process that discretizes non-measurable data (e.g., mouse age) into categorized measurement data (e.g., 1 for young mouse, 2 for middle-aged mouse, and 3 for old mouse). One approach to obtain the data items necessary for ANOVA analysis is through data item extraction from a journal article (IAO_0000443). In this case, the input is some journal article, and the output is data. The ANOVA output is a p-value data set, which includes a set of p-value results for an independent variable data set that is predefined. ANOVA is a concretization of some ANOVA protocol. The ANOVA protocol includes a predictive model that specifies a testable hypothesis model (Figure 3.3).

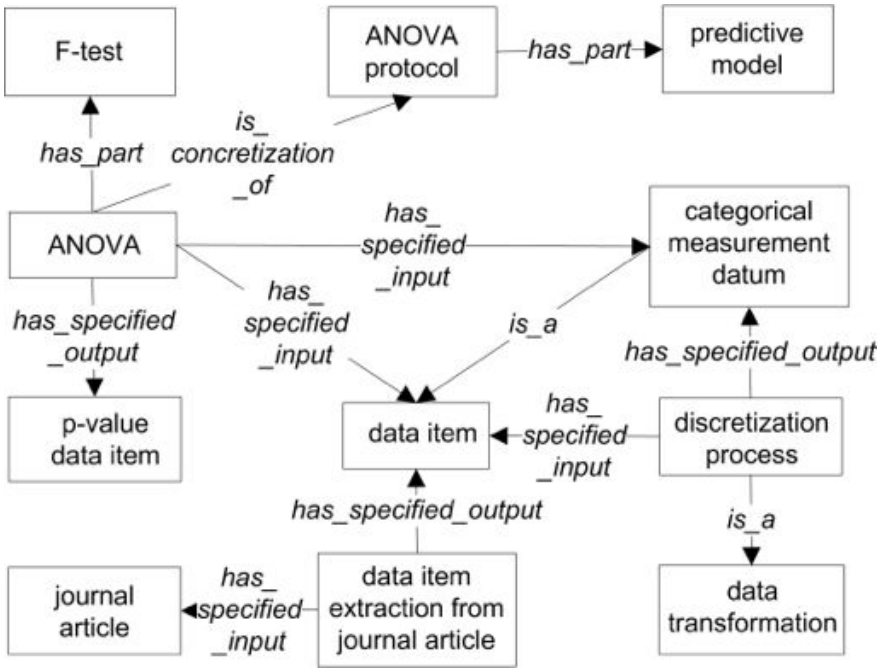


Figure 3.3: Representation of ANOVA analysis process.

Ontology representation of Brucella vaccine protection investigation

A vaccine protection investigation includes three processes (or steps): vaccination, pathogen challenge, and vaccine protection efficacy assessment. For those pathogens that kill a model animal (e.g., mouse), survival assessment is used for assessing vaccine protection efficacy [122]. Since virulent Brucella does not kill mice, the survival of pathogen challenged mice is not applicable to assess Brucella vaccine efficacy. Instead, a colony forming unit (CFU) reduction assay is used to determine the difference of live bacteria recovery from vaccinated mice and non-vaccinated mice [123]. This use case was used to derive an instance level representation based on the formal semantic representation of ANOVA analysis (Figures 3.3 and 3.4).

To determine which variables play significant roles in changing the Brucella vaccine protection efficacy, collaborators at the University of Michigan manually curated more than 40 papers to get instance data that correspond to these variables. In total, 151 instance data were collected from the literature and represented in OWL format. When variables did not already exist in the ontology they were added. An ANOVA analysis was performed and indicated that six variables do not statistically significantly contribute to the protection (p-value >0.05). These six variables include IL-12 vaccine adjuvant, mouse sex, vaccination route, mouse age at vaccination, vaccination-challenge interval, and challenge dose. The other 10 parameters statistically significantly contribute to the vaccine protection (p-value < 0.05) (Table 3.2).

Table 3.2: Ontology terms for 17 variables in the Brucella vaccine protection assay. The first variable is dependent variable, and the others are independent variables. The last six variables did not contribute to the vaccine protection (p-value < 0.05).

	Classes / ANOVA variables	Sources and term IDs
1	vaccine protection efficacy	VO: VO_0000456
2	vaccine strain	VO: VO_0001180
3	vaccine viability	VO: VO_0001139
4	vaccine protective antigen	VO: VO_0000457
5	mutated gene in vaccine strain	VO: VO_0001195
6	vaccination mouse strain	VO: VO_0001189
7	vaccination dose specification	VO: VO_0001160
8	pathogen strain for challenge	VO: VO_0001194
9	pathogen challenge (subclass)	OBI: OBI_0000712
10	CFU per volume	UO: UO_0000212
11	CFU reduction	VO: VO_0001164
12	IL-12 vaccine adjuvant	VO: VO_0001147
13	biological sex	PATO: PATO_0000047
14	vaccination (subclass)	VO: VO_0000002
15	animal age at vaccination	VO: VO_0000897
16	vaccination-challenge interval	VO: VO_0001191
17	challenge dose specification	VO: VO_0001161

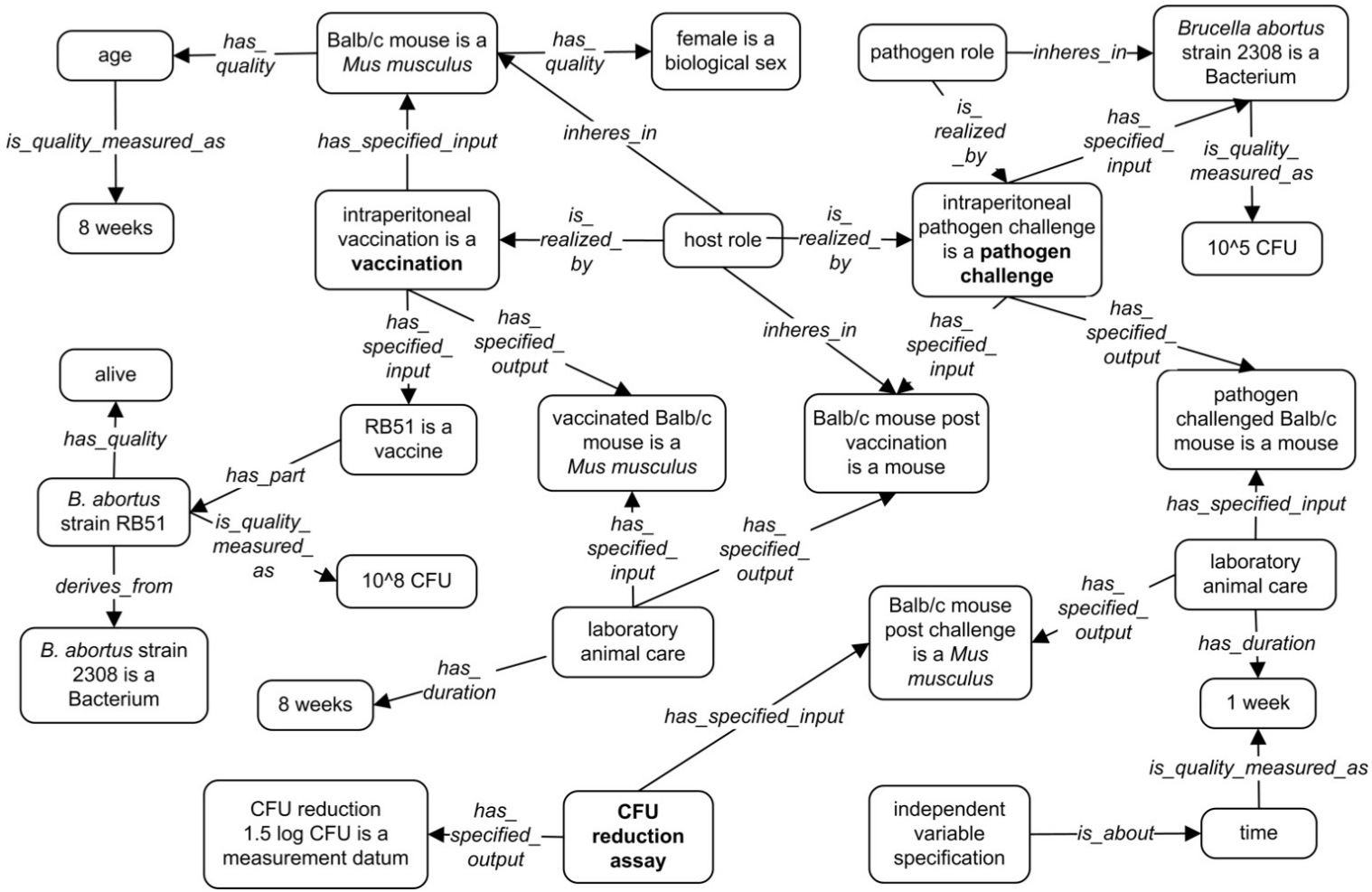


Figure 3.4: Representation of a protection assay with Brucella vaccine RB51 [123]. Boxes represent OWL individuals. Terms from different ontologies (e.g., OBI, VO, IAO) are used. Italicized text in the middle of arrows represents relations. The bold terms represent three major processes in the vaccine protection investigation

3.4 Conclusion

In this chapter, examples of how to represent experimental processes with OBI were described through three real world use cases. Experience such as this helps validate OBI's design choices, and shows how to extend it in domain specific ways. It also generates competency questions that allow us to identify parts of OBI that are insufficiently expressive and to identify external resources that can be used to extend OBI's coverage.

A major challenge when developing models is the requirement to import terms from other ontologies to construct logical definitions: due to its broad scope OBI spans multiple existing ontological resources. There is a significant cost preventing those large imports, as reasoning becomes slower and the ontology is harder to navigate. To solve this problem I developed the MIREOT mechanism [124], described in Section 6.3, which preserves namespaces of imported terms and allows their direct use into OBI and other resources.

While OBI provides a general framework for biomedical investigations, and contributed largely to other efforts such as the IAO [125] or the BFO [85], it doesn't describe clinical information such as encounters, disease and disorder processes, signs or symptoms, which are critical element when considering pharmacovigilance and adverse event reports. Those fall into the scope of the Ontology for General Medical Science (OGMS); Chapter 8 describes how AERO extends OGMS. Additionally, my work focuses on adverse events following immunization, and in that context it is highly relevant to have an accurate representation of vaccine and their components, as well as the vaccination process. Chapter 4 describes the Vaccine Ontology (VO), which targets this domain specifically.

Chapter 4

Representing vaccine data

4.1 Introduction

Vaccine research, development, testing, and clinical use involve complex processes whose computational representation requires a large number of data types and significant data volume. Several vaccine types are available; for example, live attenuated vaccines, subunit vaccines, and DNA vaccines. Vaccines are developed using multiple approaches including studies of gene and protein expression, molecular and cellular interactions, and tissue and whole body responses, as well as in extensive epidemiological modeling. Currently there are more than 200,000 vaccine-related articles in PubMed [56]. In addition to the wealth of peer-reviewed literature on vaccines, there are many public vaccine databases including the USA CDC Vaccine Information Statements system[8], the licensed vaccine information by the U.S. FDA[9]), and the Vaccine Resource Library[10]. These databases emphasize the clinical uses and regulatory oversight of existing vaccines. With the large number of vaccine data types and publications available, it is a challenge to develop an efficient strategy for vaccine data standardization, retrieval, and integration. High-throughput computational processes are needed for efficient integration of complex and large volumes of data. It is also increasingly challenging to identify and annotate vaccine data from this large and diverse literature which no one scientist or team can fully master. However, computational analysis is not possible without individual representations of various data types understandable by computers. As a result of the limited capability for data integration, efficient computational reasoning is hindered. Therefore, it was necessary to develop a common, community-supported ontology for vaccine research with both natural language and logical definitions of the terms involved.

To promote vaccine data standardization, integration, and computer-assisted reasoning, the Vaccine Ontology (VO; http://www.violinet.org/vaccineontology) was developed by the VO devel-

[8] http://www.cdc.gov/vaccines/pubs/vis/

[9] http://www.fda.gov/cber/vaccines.htm

[10] http://www.childrensvaccine.org/

opers group, of which at the time of this work Oliver He at University of Michigan, Bjoern Peters at La Jolla Institute for Allergy & Immunology, Alan Ruttenberg at University of Buffalo and myself were active. This chapter introduces the overall VO design, some core VO terms, and examples of how the VO can be used to answer specific questions in the vaccine domain.

4.2 Vaccine ontology overview

The VO was developed using OWL [114] and the Protégé editor [126]. In compliance with the OBO Foundry ID policy described in section 6.2.1, the latest version of VO is always available at `http://purl.obolibrary.org/obo/vo.owl`. In addition, VO has been deposited in the NCBO BioPortal [127], and is listed on the OBO website [55].

Most of the VO terms are for specific vaccines, indicating that the ontology is focused on the categorization and relationships of vaccines and vaccine components, vaccination investigation, and the vaccine-host interactions. Vaccine-induced immune responses and vaccine protection against targeted diseases or pathogens are derived from the fundamental vaccine-host interaction and emphasized in VO.

Some terms assigned with VO identifiers may not be vaccine specific but cannot be found in external ontologies. For example, the term 'edible' indicates the ability of a material entity (e.g., vaccine) that is orally ingestible. This term may be better located in other ontologies such as the Phenotypic Quality Ontology (PATO) [128], which focus is on those terms describing qualities of entities. In this case, a unique VO identifier has been assigned to this term for now, and submitted a term request to PATO, a typical approach to collaborative work following the OBO Foundry principles. VO is interdisciplinary and interoperable with other ontologies, especially those OBO Foundry candidate ontologies. Terms from other ontologies are imported in order to avoid duplication and support interoperability of scientific data annotated with them, data that typically spans disciplinary boundaries. VO utilizes the Basic Formal Ontology (BFO) [79] as an upper level ontology. The relation terms defined in the RO [54] have been used in VO for representing commonly used relations. VO also utilizes the IAO [125], an ontology of information entities based on the BFO. VO imports BFO, RO and IAO entirely as these ontologies are the most frequently used and their relatively small sizes don't hinder efficient editing.

However, as current editing tools fail to handle larger size ontologies, all terms from ontologies such as FMA [129] or the NCBI taxonomy [130] cannot be imported into VO. In addition, these

resources cover a broader scope that the VO, and many of their terms are not required in most cases. To import ontology terms from such large external ontologies, and prevent the need for duplication of terms already defined in other ontologies, VO relies on the MIREOT standard described in Section 6.3. OntoFox, described in Section 6.4, was used to import external ontology terms into VO. Currently, VO has imported terms from 12 external ontologies, such as the OBI [122], the Infectious Disease Ontology (IDO) [131], and the PATO [128]. For example, VO imports the term 'pathogen' (`http://purl.obolibrary.org/obo/IDO_0000528`) from IDO using OntoFox.

4.3 Specific terms defined in the vaccine ontology

4.3.1 VO definition of the term 'vaccine'

VO defines a vaccine as a processed material with the function that when administered, it prevents or ameliorates a disorder in a target organism by inducing or modifying adaptive immune responses specific to the antigens in the vaccine. In Manchester syntax [110], 'vaccine' is a defined class, i.e., translating the above constraints into logical restrictions:

```
Class: vaccine
    EquivalentTo:
        'processed material'
         and ('has function at some time' some
            ('vaccine function'
             and ('realized in' some 'vaccine immunization')))
         and (is_specified_output_of some 'vaccine preparation')

    SubClassOf:
        'processed material'
```

To translate this to prose, a vaccine is designed to perform a specific vaccine function. The 'vaccine function' can be a 'preventive vaccine function' or 'therapeutic vaccine function'. The preventive vaccine function is a vaccine function realized by the process of vaccination and leading to induction of an adaptive immune response to the antigens in a vaccine, which protects against a specific disorder, or in Manchester Syntax:

```
Class: 'preventive vaccine function'
    SubClassOf:
        'realized in' some 'disorder prevention',
        'realized in' some 'induction of adaptive immune response to antigen',
        'vaccine function',
        'realized in' some vaccination
```

The 'therapeutic vaccine function' is defined similarly. Correspondingly, there are two types of vaccines: preventive vaccine and therapeutic vaccine. According to the Ontology for General Medical Science (OGMS [132]), a disorder is the physical basis of a disease such as infectious disease, cancer, allergy, or autoimmune disease. VO uses these disorders to build its asserted structure and define vaccines. For example, the class 'human immunodeficiency virus vaccine' is defined as a viral vaccine that is administered to prevent an infection of human immunodeficiency virus, or as represented using the Manchester syntax:

```
'viral vaccine'
and administered_to_prevent some (infection_of some 'Human immunodeficiency virus')
```

At the University of Michigan, VO was used to develop VIOLIN (`http://www.violinet.org`), a web-based vaccine database and analysis system to store and analyze research data concerning commercial vaccines and vaccines under clinical trials or in early stages of development [121]. Based on VIOLIN and users requirements, 301 vaccines or vaccine candidates for 20 different genera or species of animals have been included in VO, including all 146 vaccines licensed for human use in the USA and Canada. Many of these vaccines are also used in other countries. More efforts are under way to include additional licensed vaccines in VO.

Vaccines can also be classified depending on the vaccine preparation method, such as 'inactivated vaccine' and 'subunit vaccine'. To facilitate research and development of these different vaccines, these terms have been included. However, multiple inheritance (i.e., a child term linked with multiple parent terms) classes will occur if a vaccine is classified under a vaccine that induces immunity in vivo against infection of a pathogen (e.g., Influenza virus) and a vaccine that is prepared by inactivation of the whole pathogen and using the inactivated pathogen as vaccine antigen. For example, a vaccine (e.g., Afluria) may be categorized as both an 'Influenza virus

vaccine' and an 'inactivated vaccine'. To increase explicitness, modularity, and maintainability, asserted multiple inheritance should be avoided during ontology development [133]. To address this, in VO, the asserted hierarchy is based on the OGMS disorder hierarchy and OWL reasoners are used to infer additional information. For example, Afluria is asserted under Influenza virus vaccine. The Afluria vaccine antigen is the whole viral organism that has quality "inactivated". The Afluria vaccine is therefore declared as bearing the quality "vaccine organism inactivated". As "inactivated vaccine" (`http://purl.obolibrary.org/obo/VO_0000315`) is defined as:

```
EquivalentTo:
    vaccine
     and ('has quality at all times' some 'vaccine organism inactivated')
```

a reasoner will classify Afluria correctly as an "inactivated vaccine".

4.3.2 VO definition of the term 'vaccination'

The term 'vaccination' is another core term in VO. Vaccination is the process of administering a vaccine into an organism (e.g., human). The definition of this VO term relies on three OBI terms. Specifically, vaccination (VO_0000002) is modeled as a process of 'administering substance in vivo' (OBI_0600007), in which some 'vaccine' realizes the 'material to be added role' (OBI_0000319) to an organism (OBI_0100026):

```
'administering substance in vivo'
and realizes some ('material to be added role' and role_of some vaccine)
and realizes some ('target of material addition role' and role_of some organism)
```

Figure 4.1 is an example of 'vaccination' with Afluria influenza vaccine. Specifically, the Afluria vaccine which bears the 'material to be added role' is administered in vivo into a mouse ('target of material addition role'). The vaccine is contained in a vial ('containing function') and drawn into a syringe ('injection function') for vaccine injection. This vaccination is implemented with an administration dose of 0.2 ml ('administration dose role') and through the intramuscular route ('administration route role'). As a result of this work, the whole process of administering Afluria can be described in an OWL file allowing computers to understand and parse a vaccination process, and thus support automated reasoning.

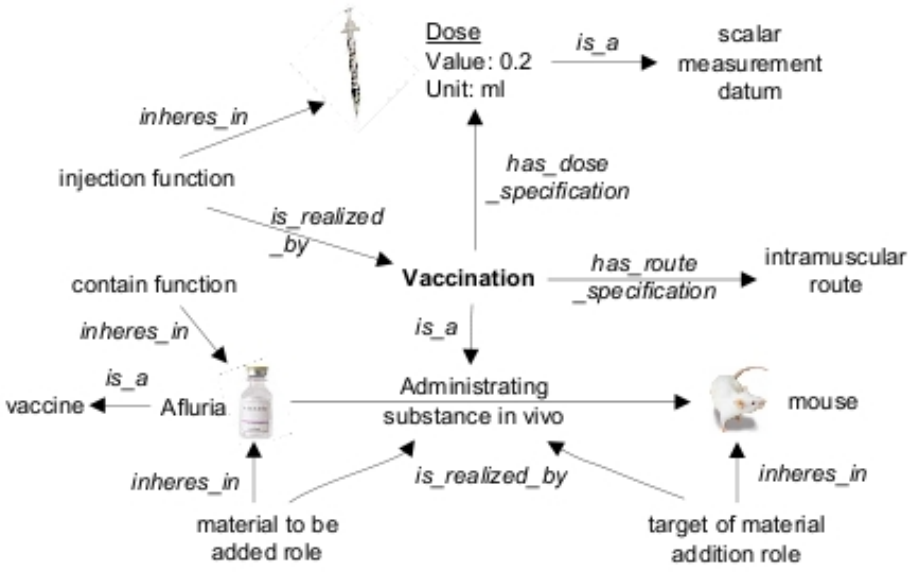


Figure 4.1: Representation of vaccination (VO_0000002) using VO and OBI. All relation terms are italicized.

4.3.3 VO representation of immune response to a vaccine

The study of immune responses in an organism administered with a vaccine is critical to vaccine research and development. The classes under the VO term 'vaccine-induced host immune response' are shown in Figure 4.2. Those immune responses important for protection against various diseases are emphasized. Specifically, vaccine induces adaptive immune responses including immunities mediated by B cells or T cells, and T helper type 1 or 2 immune responses. Antigen processing and presentation is undertaken in B cells and other professional antigen presenting cells (e.g. macrophages and dendritic cells). B cells give rise to antibody-mediated immune responses, while T cells give rise to cytotoxic T lymphocyte activities. A T helper 1 (Th1) type immune response is normally required to protect against infections caused by viruses (e.g., Poliovirus) and intracellular bacteria (e.g., Brucella), while a T helper 2 (Th2) type immune response is usually required to protect against extracellular bacteria (e.g., E. coli). Meanwhile, a vaccine also induces activation of various cells including dendritic cells and lymphocytes. The above information has been included in VO (Figure 4.2). In addition, VO includes information about vaccine-induced innate immune responses, which are often stimulated by vaccine components (e.g., adjuvant [134]). Vaccine-induced activation of various cell types is also included in VO (Figure 4.2).

Although not explicitly stated, the current VO terms of vaccine-induced immune responses

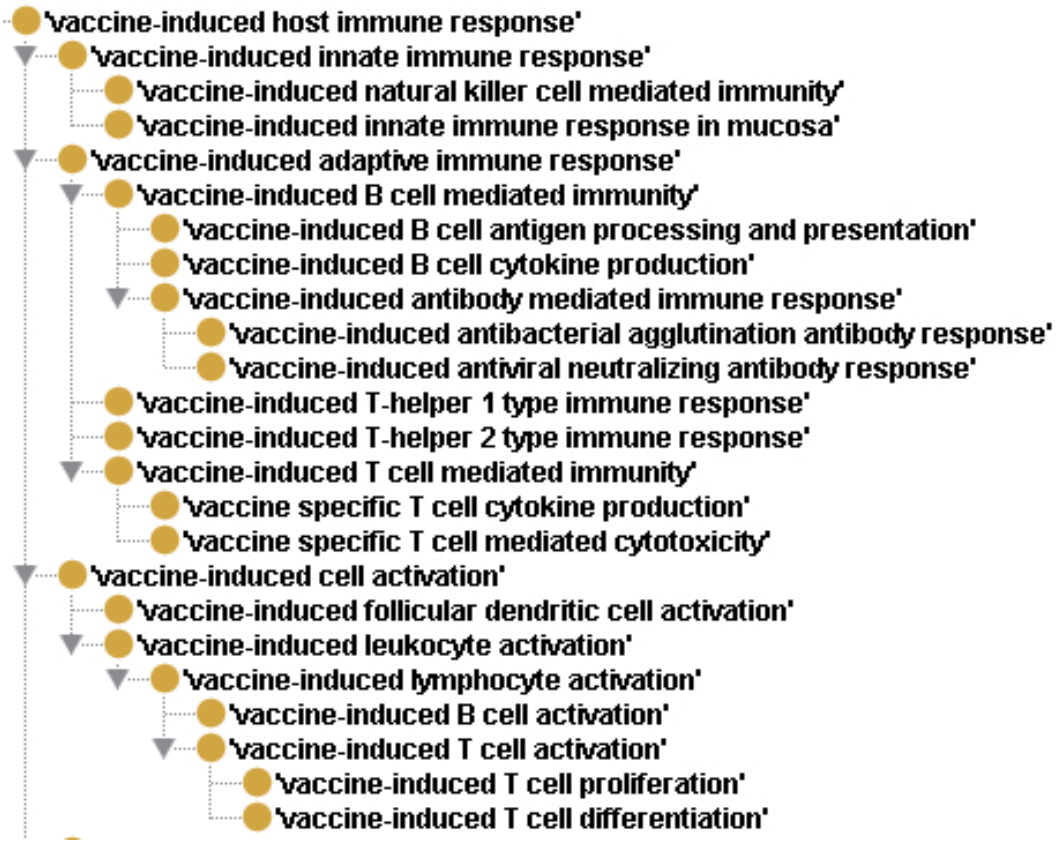


Figure 4.2: Hierarchy of vaccine-induced immune response in VO.

reference corresponding immune responses introduced in the Gene Ontology (GO) [53]. A vaccine-induced immune response (e.g., vaccine-induced T-helper 1 type immune response) can be considered as a cross product between a corresponding GO term (e.g., T-helper 1 type immune response) and a VO-specific term (e.g., vaccine-induced adaptive immune response). The GO term 'T-helper 1 type immune response' (GO:0042088) is associated with 219 gene products (`http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0042088`, [135]). Further investigations are required to determine whether all or a portion of these 219 genes products are indeed associated with a vaccine-induced T-helper 1 type immune response.

4.4 Vaccine ontology applications

4.4.1 Naming vaccine-specific terms

VO contains different aspects of vaccine composition and biology and can, therefore, be used to model individual vaccines. An example of modeling two influenza vaccines, Afluria (`http://www.`

afluria.com/) and FluMist (http://www.flumist.com/), is illustrated in Figure 4.3. Afluria is an inactivated influenza vaccine manufactured by CSL Limited and administered intramuscularly. FluMist is a live attenuated influenza vaccine manufactured by MedImmune and is administered intranasally. Both Afluria and FluMist share many similar allergens (e.g., chicken egg protein). Due to their different vaccination routes, different types of adverse events may be induced. For example, Afluria induces injection-site pain and muscle ache, while FluMist induces cough and sore throat. The similarities and differences shown in Figure 4.3A can also be transferred into the computer-readable ontological representation (Figure 4.3B).

This model contains many terms unique to VO, such as vaccine, influenza vaccine, and the names of these two vaccines, all of which have been assigned VO specific identifiers in the VO namespace. This model also contains many terms that originate from other ontologies. For example, Influenza virus A and B are imported from the NCBI taxonomy [130] using the MIREOT [124] system and retain their original identifiers. This approach allows VO to maintain an optimized structure for modeling vaccine-specific features while integrating with existing ontological resources, thus ensuring orthogonality and synergy of different ontologies. The represented ontology terms and computer-interpretable format can further be used for development of different computational tools for automated reasoning. Conversely, other biomedical ontologies such as OBI [122] and the Influenza Ontology (InfluenzO [136]) import specific VO terms as part of their development process.

4.4.2 Vaccine data exchange and integration

The VO is loaded on the Ontobee server (described in Chapter 7), and can be queried via the corresponding SPARQL endpoint, at http://www.ontobee.org/sparql/index.php. For example, the SPARQL queries shown in Appendix D retrieves all information about the FluMist vaccine (VO_0000044). SPARQL queries can also span multiple biomedical ontologies, allowing for integration with other resources.

4.4.3 Development of vaccine knowledgebase and semantic web

A VO-based vaccine knowledgebase can be generated by representing the data curated in the VIOLIN vaccine database [121] as instances of VO in a standardized approach to comply with the VO requirements. VIOLIN contains a large amount of data about vaccine protection experiments. All VIOLIN vaccine protection data can be represented as VO instances using the OWL format by using and expanding the VO modeling of vaccine protection assays. Instances of vaccine protection

assays using the data from the VIOLIN database have been generated. Such an integration approach allows users to query vaccine protection experimental data using complex SPARQL queries and applying the results for advanced vaccine analysis.

4.4.4 VO-based literature mining

VO can be used to facilitate vaccine literature mining. Progress in vaccine research has led to a dramatic increase in the number of vaccine-related papers. As a result, it has become increasingly challenging to retrieve relevant vaccine data for research purposes. There are currently more than 200,000 vaccine-related journal publications based on a search of "vaccine OR vaccination" in the PubMed literature database [56]. PubMed articles are annotated with the Medical Subject Headings (MeSH, [137]). However, MeSH contains limited vaccine-specific information. For example, Brucella is an intracellular bacterium that causes brucellosis, the most common zoonotic disease worldwide [138]. MeSH contains the term Brucella vaccine but does not include any subclasses under this term, limiting the search for Brucella vaccine in PubMed. However, 40 specific Brucella vaccines are currently represented in VO as subclasses (children) of the VO term Brucella vaccine. Each subclass in VO has an is_a relationship with its parent class. This ensures that all subclasses (e.g., Brucella RB51) can be included when a parent class (e.g., "Brucella vaccine") is searched. Inclusion of these 40 specific Brucella vaccines and their synonyms as keywords in PubMed searching Brucella vaccine increased the search results by 25% from 1296 to 1619 (as of September 19, 2009). Specific annotations of different vaccines in VO can also be used for literature searching.

A user case study is to search for "live attenuated Brucella vaccine" in PubMed. As of June 16, 2009, a direct PubMed search of this string of keywords returned 58 papers (or PubMed hits).

A search using VO information, performed at University of Michigan, dramatically increased the recall of searching "live attenuated Brucella vaccine" by 13 fold (693/55) compared to the searching without using VO (Table 4.1).

Those results also showed that the precision of the searching remains high (96%), demonstrating that VO can be used to significantly improve PubMed searching efficacy in the vaccine domain.

4.5 Discussion

As a collaborative community-based effort, VO is closely related to many other biomedical ontologies. As vaccines are integral in the prevention of many infectious diseases, VO has strong ties

Table 4.1: VO enhanced literature search.

PubMed Search Keywords	Hits	True	Precision
live attenuated Brucella vaccine	58	55	95%
Consider live attenuated Brucella vaccine in VO:			
Brucella (RB51 OR SRB51)	182	182	100%
Brucella (strain 19 OR S19)	537	510	95%
Brucella Rev.	145	144	99%
B. suis (strain 2 OR S2)	11	10	91%
Brucella bacA mutant vaccine	1	1	100%
Other 12 live attenuated Brucella vaccines in VO	62	59	95%
Total (unique ones)	720	693	96%

with IDO. VO encompasses vaccines against various vaccine-preventable diseases with a particular emphasis on infectious diseases. Since a vaccine can be developed against different stages of the life cycle of an infectious pathogen, the combined application of VO and IDO will provide a superior means for analyzing differing vaccine strategies. VO development has been and continues to be closely related to the development of the OBI (described in Chapter 3. Many VO terms (e.g., vaccination) pertain to various vaccine experiments that are in the purview of OBI. Continued close collaborations between these projects will ensure coordinated evolution of the many different resources.

Some key challenges remain for future development in VO and sister ontologies such as IDO, OGMS and OBI. For example, the relations among disease/disorder, organism, and infectious disposition are currently under debate. In the current version of VO many new relation terms have been defined, such as "administered_to_prevent" and "has_route_specification". However, using the RO defined relations would allow easier querying across different ontologies, and provide a clearer understanding of relations between entities [54]. A common representation between those efforts is therefore preferable. The new relation terms are currently defined in VO as an intermediate step in VO development, and will be reexamined in future work, looking for opportunities to reuse existing

relations defined in RO, in collaboration with the RO developers.

Another challenge is that those ontologies whose terms are often needed for VO development are still under development and extension. For example, extensions are needed for the NCBI taxonomy. An infection of human by an influenza virus is a disorder that forms the physical basis of a human influenza disease. In the NCBI Taxonomy, Influenzavirus A-C are three genera under the Orthomyxoviridae family. 'Unidentified influenza virus' is one species under unclassified Orthomyxoviridae. There is no single term called Influenzavirus that covers all these different Influenza viruses, which all may cause an influenza disease. Currently there are 20 licensed Influenza vaccines stored in VO. In most cases, each Influenza vaccine covers more than one Influenza genera or species. To simplify the description, a new term Influenzavirus may need to be generated in NCBI Taxonomy in order to cover these four types of influenza viruses. In this case, a suggestion to the NCBI Taxonomy team for inclusion of the new term Influenzavirus should be submitted. In summary, a collaborative effort is required to progress VO and sister ontologies to address different needs and challenges.

4.6 Conclusion

VO is targeted to include all licensed vaccines in different countries and regions as well as vaccines in clinical trials and undergoing development in research laboratories. This inclusion will allow advanced integration and intelligent analysis of the large amount of vaccine data produced around the world. Continuing development of VO will include additional information in such aspects of vaccines as vaccine clinical trials and vaccine surveillance. By structuring complex vaccine data types and data volumes, this approach will promote a shared understanding of vaccines.

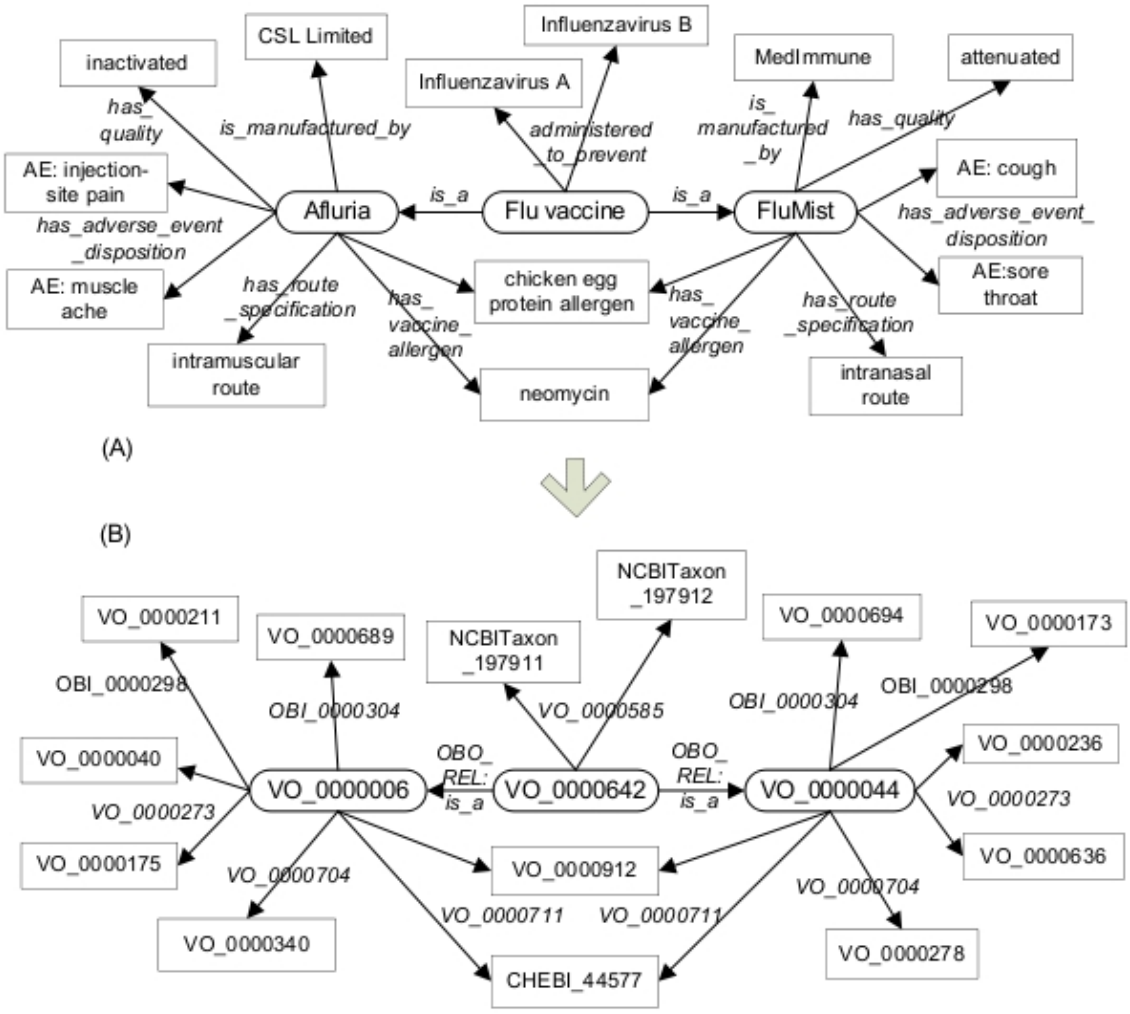


Figure 4.3: Comparison of Afluria and FluMist influenza vaccines using VO. The labels of ontology terms are shown in (A). The same content can also be represented by ontology identifiers understandable by computer programs (B). Each arrow represents a direction of a relation between two classes shown in boxes. All relations are italicized.

Chapter 5

Semi-automated ontology building using design patterns

5.1 Introduction

A complex, expressive and logically rigorous, domain representation that can be practically maintained and validated by reasoners, such as Pellet [60] and FaCT++ [102] can be constructed through the creation of OWL [92] classes with logically necessary and sufficient definitions. However, manually adding classes and logical axioms is a time-consuming [139], possibly error prone process, in spite of using advanced ontology editors such as Protégé [126]. Also, using such editors requires rather extensive knowledge of OWL. This requirement significantly limits the number of people who can contribute productively to enriching the ontology. Considering that each year many hundreds of candidate terms are being submitted to large resources such as OBI, the process of defining them must not become a bottleneck.

The approach described here is motivated by the observation that definitions of a significant proportion of term requests can be accommodated by a limited number of pre-defined design patterns. It falls into the realm of ontology design patterns (ODP), which cover those techniques used to solve common and recurring representation problems [140, 141, 142, 143]. Relying on such ODP, a practical solution, geared toward bioontology developers and editors, has been developed. In order to engage domain experts without extensive practice in ontology development, the required input for each such design pattern as a QTT, which can be edited as an Excel spreadsheet, was formulated. Excel spreadsheet format was chosen as being the most ubiquitous, familiar, and easy to use to scientists. The work on QTT was led by Philippe Rocca-Serra, and I participated in the implementation and testing. In the following, an example of a common term request is illustrated, namely assays that measure the concentration of a specified molecular compound in a given material, which is typical for clinical chemistry assays. Requests for terms to identify

such assays come from diverse communities, including EBI's BioInvestigation Index [144, 145], the Immune Epitope Database [146] and the Influenza Virus BioHealthbase [147]. This example illustrates the QTT process as a proof of principle.

5.2 Methodology and results

The Quick Term Template submission process has four main steps as shown on Figure 5.1:

1. Agreement by the OBI consortium on the logical definition of the parent class for submissions matching a certain pattern;

2. Identification of entities that can be varied with respect to the parent class (the differentia), for which a QTT spreadsheet containing one column for each such entity is generated;

3. Population of a QTT spreadsheet by domain experts;

4. Processing the QTT submission to generate new classes and definitions, and returning the label and identifier of an OBI class for each valid row of the submission.

5.2.1 Step 1: Develop the representation of the parent class

The example used throughout this section is a QTT for subclasses of 'analyte assay' in OBI (OBI:0000443). Such assays measure the concentration of a specified molecular entity relative to a given material entity, such as measuring glucose concentration in blood in units of μg per liter. Each logical definition relates the material in which the concentration is measured (the evaluant; e.g., blood), the molecular entity that is detected (the analyte; e.g., glucose), and the units of the measurement being made (e.g., microgram per liter). The full logical definition of this class is shown in Figure 5.2: an analyte assay achieves planned objective some analyte measurement objective. During the analyte assay, the evaluate role is realized (e.g. by the blood) and the analyte role is borne by some scattered aggregate constituted how homogeneous grains (e.g., glucose in the blood). The output of the assay is information about concentration - a relational quality of the analyte towards the evaluant. Figure 5.3 shows how the analyte measuring assay is modeled in OBI. The corresponding textual definition is: "An analyte assay is an assay with the objective to determine the concentration of one substance (bearer of the analyte role) that is present in (part of) another (bearer of the evaluant role)".

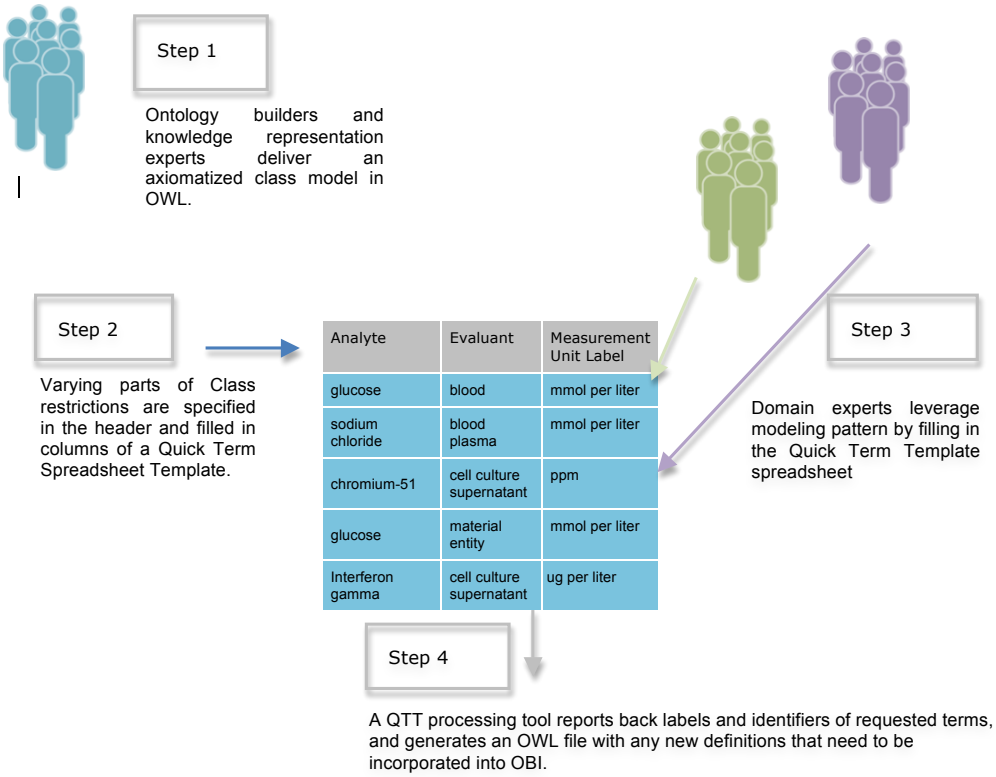


Figure 5.1: Overview of the process using the OBI modeling of the class 'analyte assay' as a starting point or seed class (Step 1). The variable parts (differentiae) of the representation are used to derive a Quick Term Template made up of 3 fields (Steps 2 & 3). The OWL file is generated using a dedicated tool (Step 4). For simplicity, identifiers were omitted in this figure.

5.2.2 Step 2: Derive tabular Quick Term Template

A large number of current requests for terms are subclasses of analyte assay. Their differentiae are the analyte (i.e., what the concentration is being detected of), the evaluant (i.e., the material in which the analyte concentration is detected) and the unit, which is used to qualify the measurement datum. Consequently, a Quick Term Template for an analyte assay needs columns for only those three entities. Table 3.1 depicts a QTT with several example entities, as they would be seen in a spreadsheet by a submitter. Each column is to be filled with elements that are of a specified general type. The analyte column is expected to be a subclass of molecular entity, and the evaluant column any material entity.

```
Analyte assay:
    achieves planned objective some analyte measurement objective
    and realizes some ('evaluant role' and (role of some material entity))
    and realizes some ('analyte role' and
                                    ( role of some ('scattered molecular aggregate' and
                                                    has grain  only  molecular entity)))
   and has_specified_output
some ('scalar measurement datum'
              and ('is quality measurement of' some 'molecular concentration')
              and ('has measurement unit label' some concentration unit label'))
```

Figure 5.2: OWL restrictions that logically define the analyte assay class in OBI.

5.2.3 Step 3: Domain experts populate the template

The template hides the complexity of modeling by only identifying the differentiating entities needed for the definition of the class while hiding the actual relations binding those entities together. The burden of adding logical definitions is displaced advantageously from the users to the machine, which reliably and automatically populates class specifications from the template and the differentiating entities supplied by the user. A template such as this one would be accompanied by guidelines for users explaining what values are allowed in the columns, and how they will be interpreted in building the assay.

5.2.4 Step 4: Submission processing

Following submission of a completed QTT, a QTT processing goes through the steps outlined below.

1. Identify referenced classes from external ontologies, and import them as necessary via the MIREOT mechanism [124]. OBI relies on this mechanism to reference classes in external ontologies. In case entities are absent from resources, a submission is necessary. Request processing was quick enough not to be perceived as a hindrance to the process.

2. Create an OWL class description by substituting values from the spreadsheet.

3. Use the constructed class description to do a query for equivalent classes already in OBI. If

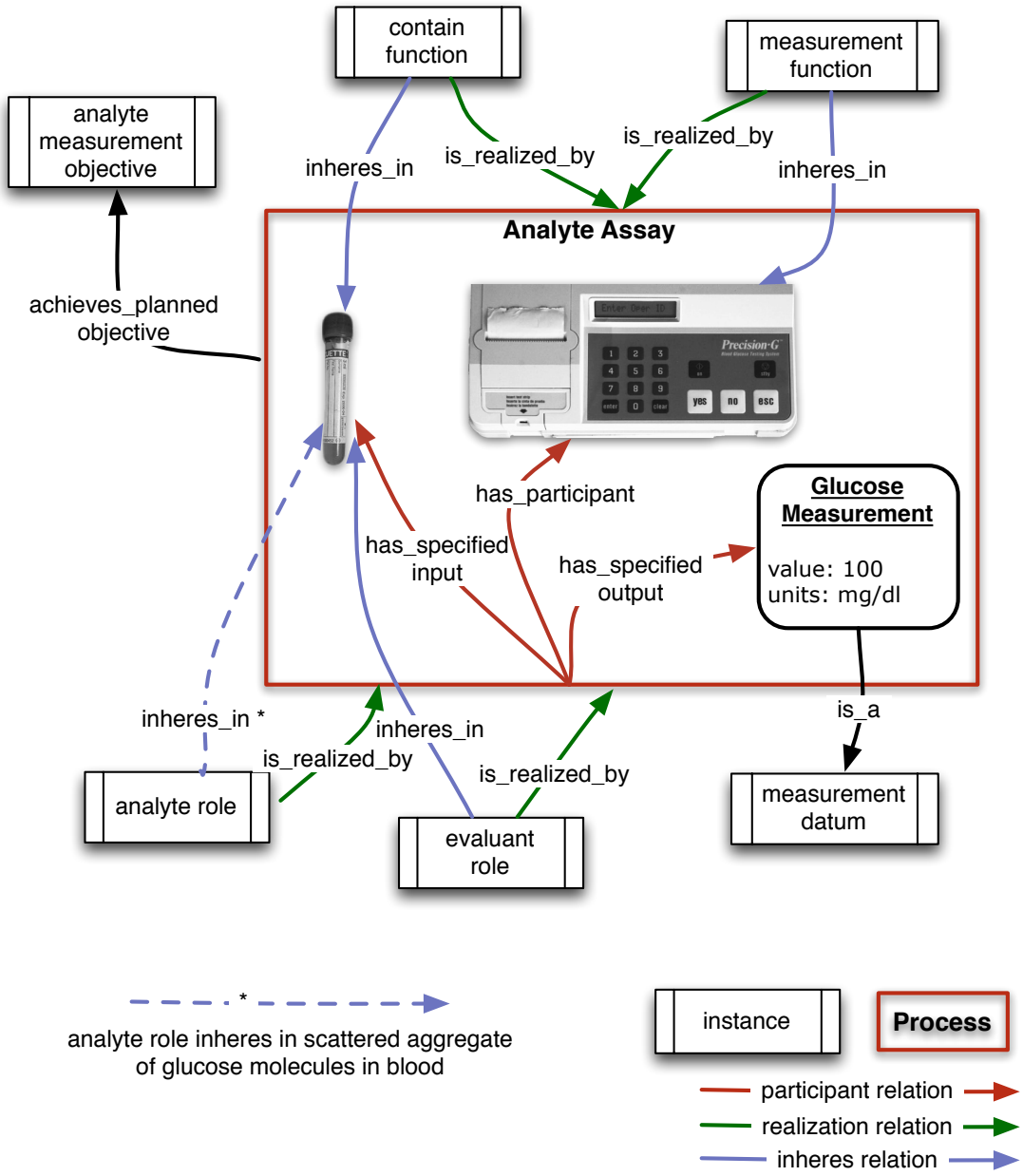


Figure 5.3: Analyte assay class in OBI. Adapted with permission from Alan Ruttenberg.

an equivalent class exists, store its URI and label and continue to the next row.

4. If an equivalent class does not already exist in OBI, create a unique OBI URI and associate with it a new class defined by the constructed class description. Add metadata such as label and definition. As a QTT submission creates fully logically defined classes, the creation of labels and textual definitions can be automated. For the examples in Table 3.1, the class defined by Row 1 is assigned the label 'glucose concentration measurement in blood in units

of mmol per liter', the class corresponding to Row 5 the label 'interferon gamma concentration measurement in cell culture supernatant in units of μg per liter'.

5. Use a reasoner to perform a consistency check and classification of the combination of the existing OBI file and one with the newly created classes.

6. Report on processing of the template, and return a list of URIs and labels corresponding to the rows of the submitted QTT spreadsheet.

5.3 Implementation

The implementation of steps 1-3 in the QTT approach requires no automation as their focus is on providing the template to be used throughout. The 'analyte assay' example is shown in Figure 5.2 and Table 5.1. Step 4 requires implementation of an automated QTT handler. In order to validate the approach, a prototype of a QTT handler was created using Perl. Plain OWL templates were derived from the previously described class representation as found in the ontology and populated with token values parsed out from the incoming QTT spreadsheet template (Step 4.5). This prototype implementation delivered the expected results and helped in refining requirements of the workflow for Step 4. However, running this implementation required extensive manual intervention making it not end-user friendly. Therefore other options were considered.

As OBI developers are heavy users of the Protégé editor, its plugin library was explored for alternative implementation options. Three add-ons (the Matrix (Drummond, 2008), Excel Import [148] and OPPL [143] plugins for Protégé 4) were evaluated. The Matrix plugin enables tabular visualization of the axioms of an existing class viewed in Protégé. Potentially, this allows rapid crafting of a Quick Term Template from a class; however, the lack of a persistence mechanism for saving such a template significantly limits direct applicability of the Matrix plugin for the QTT approach. Excel Import plugin has a fairly explicit aim: taking in an Excel spreadsheet and creating OWL classes by relying on a set of rules to declare the relations between columns. This is technically very close to what is required for implementing the QTT specification. However, while assessing the relevance of the Excel Import plugin, two major stumbling blocks appeared. First, the incoming spreadsheet had to be explicit, meaning that all restrictions and fillers placeholders must be present as column headers for the OWL generation to occur properly. This requirement defeats the purpose of the QTT, which aims to conceal some of the modeling the complexity from end users. Second, it is not possible to create nested axioms on a class X such as 'X (realizes some ('evaluant

role' and (role_of some Y))) from the tool's Restriction Generator pane which doesn't provide such option. This second limitation means that Excel Import could be used for fairly simple and direct class restrictions but is incompatible with some of the more advanced patterns being tested by OBI developers.

More flexibility in specifying axioms and manipulating OWL ontologies is provided by the OPPL plugin, which relies on the Manchester syntax [110]. However, the OPPL plugin requires Protégé 4 while some of OBI development still relies on Protégé 3.4 features. All three of these tools are highly useful and fully functional for their intended applications, but each was missing functionality necessary to develop an end-to-end prototype for QTT template processing, as detailed above. Specifically, none provided the ability to build class expression templates of the complexity shown in Figure 5.2, persist such templates, and populate them by parsing information from a spreadsheet.

Therefore the process was designed to use a prerelease version of MappingMaster, a plugin for Protégé 3.4 [149] for mapping spreadsheet content into OWL that is under active development. The following section describes experience with this tool. It provides a Domain Specific Language (DSL) that is based on the Manchester Syntax to define these mappings. In this DSL, any reference to an OWL named class, OWL property, OWL individual, or a data value can be substituted with a reference to one or more cells within a spreadsheet. Any expressions containing such references are preprocessed and the relevant spreadsheet content specified by these references is imported. This content can then be used in four main ways:

(1) It can be used to directly name OWL entities that are created on demand.

(2) It can be used to annotate OWL entities that are created on demand.

(3) The content may reference existing OWL entities, either directly as a URI or through an annotation.

(4) The content may be used as a literal data value.

Using one of these approaches, each reference within an expression is thus resolved during preprocessing to a named OWL entity or a data value and the resolved value is substituted for its associated reference. A standard Manchester syntax processor (e.g., the OWLAPI [150]) can then interpret the resulting expression and generate the OWL equivalent statement. Declaratively specifying mappings in this way has several advantages. No programming or scripting expertise is

required to write those mappings, and they can be easily shared using the MappingMaster plugin where they can be persistently stored as OWL files. The mappings can then easily be executed repeatedly on different spreadsheets with the same structure. Since MappingMaster is available as a Protégé plugin, the results of mapping processing can be examined immediately within the ontology editor, and the mappings modified as needed and immediately re-executed, speeding the development process. MappingMaster also includes an interactive editor for the mapping DSL that supports on-the-fly entity name checking and dynamic expansion of entity references.

The DSL expressions used to convert the QTT template into OWL classes as shown in Figure 5.4 are passed to MappingMaster. Running the tool generates an additional OWL file that contains all newly created classes. All the material is available from the OBI wiki [151].

```
Class: @A*(rdfs:label 'analyte assay')
EquivalentTo:
(achieves_planned_objective some 'analyte measurement objective') and
(realizes some ('evaluant role' and (role_of some @D*(material_entity)))) and
(realizes some ('analyte role' and
                (role_of some ('scattered molecular aggregate' and
                                ('has grain' only @B*('molecular entity'))))))
SubClassOf:
 has_specified_output some
     ('scalar measurement datum' and
        ('is quality measurement of' some 'molecular concentration') and
('has measurement unit label' some @F*('measurement unit label')))
```

Figure 5.4: Template expressions in MappingMaster's DSL based on the Manchester Syntax. References to spreadsheet cells are prefixed with "@". Cell values are substituted into the template by MappingMaster to generate class descriptions associated with the QTT.

5.4 Conclusion

The QTT process outlined here provides two benefits. First it provides a method to incorporate a large number of classes considered of high value by domain experts in the communities OBI is designed to serve. For example, the IUPAC clinical chemistry resources [152] contain hundreds

of assays describing analyte measurements. Second, the approach allows domain experts to directly populate templates without having to learn OWL syntax. The evaluation of MappingMaster Protégé plug-in as a QTT handler produced encouraging results. This early version already exhibits key features such as flexible creation of axioms thanks to a domain specific language based on the Manchester Syntax, as well as capabilities to automatically generate names for the newly created defined classes by passing user defined expression to the rdfs:label field. Finally, it tries to avoid class duplication by inspecting the target ontology for existing entries matching the input received from the template. In mid 2010, Carlo Torniai at Oregon Health & Science University successfully used QTT to add 100+ instruments from the Eagle-i [153] project to OBI.

As always with 'off the shelf' solutions such as MappingMaster, there are some caveats. Portions of the QTT specifications are not entirely supported and so reaching production grade reliability will require further work. Several rounds of evaluation have led to a number of feature requests that would facilitate performing the entire procedure. In particular, three areas would benefit from such efforts:

- First, it is desired to have the capability to perform automatic class resolution when dealing with external ontologies. This would require implementing the MIREOT mechanism. At present, external reference resolution in a QTT template is done manually prior to running MappingMaster, with processing aborted if any term is not found. It should be noted that ISACreator [145, 154], a spreadsheet editor geared towards managing experimental metadata ships with embedded ontology lookup service. It could be harnessed to create and populate Quick Term Template in order to address this limitation.

- Second, development could be made more efficient by checking class membership and equivalence as run-time queries rather than relying on full reclassification of the ontology, which can be very time consuming. In the analyte assays example presented, this would allow quick detection of classes declared as analyte (first column in Table 3.1) but which are not subtypes of 'molecular_entity', as expected. Reporting such errors immediately would speed up debugging of submitted QTT spreadsheets.

- Third, adding class metadata should be better supported. It is currently not possible to create class annotations such as cross-references, editor notes or alternative names, which would be easily supplied when creating the QTT spreadsheet. Future releases of MappingMaster plugin will cater for this need.

Table 5.1: A basic QTT for submitting an analyte assay term request. This specification includes classes defined in several ontologies: Chemical Entities of Biological Interest (ChEBI) [155], The Foundational Model of Anatomy (FMA) [129], the Unit Ontology (UO) [156], the Protein Ontology (PRO) [157] and BFO.

Analyte label	Analyte ID	Evaluant label	Evaluant ID	Measurement unit label	Measurement unit ID
glucose	CHEBI:17234	blood	FMA:670	mmol per liter	UO:0000300
sodium chloride	CHEBI:26710	blood plasma	OBI:0100016	mmol per liter	UO:0000300
chromium-51	CHEBI:50076	cell culture supernatant	OBI:1000023	ppm	UO:0000169
glucose	CHEBI:17234	material_entity	BFO:MaterialEntity	mmol per liter	UO:0000300
interferon gamma	PRO:000000017	cell culture supernatant	OBI:1000023	μg per liter	UO:0000301

Since this initial implementation of the QTT, a new version of the Protégé editor, Protégé 4, was released. Unfortunately, the MappingMaster plugin was not ported to this new version. However, a web-based server, the Ontorat [158] was created to allow easy incorporation of multiple terms in resources following the QTT guideline. It takes one Excel spreadsheet as input, and returns the spreadsheet containing the IDs of the terms, as well as an OWL file that can either be copied into the source file or simply imported. Some features are still missing from the Ontorat. For example, where MappingMaster could create new terms that did not exist in the ontology, Ontorat requires that all terms used already have an associated URI. Also, while Ontorat allows adding annotations on terms (or editing existing ones), it does not currently handle instances, which means some of those, such as the curation status annotation described in Section 6.2.2, are not supported. Ontorat developers are actively working towards addressing those issues.

Chapter 6

Working with large biomedical resources

6.1 Introduction

In this chapter, I describe my investigation of what elements are required to support a large consortium of ontology developers developing compatible resources for publication on the Semantic Web. After building the biomedical resources described in Chapters 3 and 4, there remains a need to need to address how they can be used together, and in conjunction with other relevant resources, specifically in the framework of the OBO Foundry described in Section 2.4. The ability to work with multiple resources is critical to allow developers to concentrate on new requirements rather than duplicate existing efforts.

In order to harmoniously build on several distinct bodies of work originating from different communities, guidelines should be established and followed. While several OBO Foundry principles already are adopted or are under discussion (see Section 2.4), critical gaps remained to be filled. To address some of those, several OBO policies were developed and are presented in Section 6.2. For example, the adoption of a common ID policy is crucial to fulfill the Semantic Web requirement of using URIs as identifiers, which will be critical for publishing resources as shown in Chapter 7. Additionally, sharing a common metadata set, as described in Section 6.2.2, not only provides a reliable, consistent behavior to the end user, but also allows building tools which support multiple resources consistently, such as is the case with MIREOT (see Section 6.3), OntoFox (see Section 6.4), or Ontobee (see Chapter 7).

One of the core principles of the OBO Foundry is to maintain orthogonality between resources: no resource should duplicate work done by another, to prevent heterogeneity in representation of entities and duplication of effort. In this context, it was critical that a mechanism be devised to allow select usage of specific classes or portions from external resources - the MIREOT. MIREOT,

described in Section 6.3, allows integration of multiple ontologies and taxonomies without being hindered by the increasing size of the result. However to be useful to the community, MIREOT needs to be easily available and implemented such that non-computer specialist can use the system. To that effect, a web-based tool that implements the MIREOT guideline in a user-friendly way was created: Ontofox, described in Section 6.4.

6.2 OBO Policies

6.2.1 Common unique identifier policy

The OBO foundry currently hosts resources under the OBO [159] and OWL [160] formats, and aims at providing tools such as the OWLAPI mapping for OBO format [11] to allow their interconversion. In order to do so, one key requirement is to rely on a common system to handle unique identifiers for entities. A policy, normative for Foundry resources, includes a Foundry-compliant URI scheme, and rules to map from current OBO IDs and OBO legacy URIs towards them. In collaboration with Alan Ruttenberg and Chris Mungall, I devised an ID policy for OBO resources.

Following a common ID policy allows URIs to be more reliable, and ensures they are unique within the Foundry consortium. It also helps building tools relying on this ID scheme. For example, the OBI [161] developers do not deal with ID management when creating entities; rather a script is run pre-release to check and homogenize URIs for format and stability (e.g., was any URI deleted since the last release?). Another feature is to allow dereferencing and provide useful information to a user trying to resolve terms' URIs. The Ontobee browser [12], described in Chapter 7, displays an HTML page that provides human readable information on each term, such as label and textual definition, while the page source is RDF that can be machine-processed. Finally, the ID policy specifies versioning rules for ontology releases, effectively creating a version history for resources. By doing so, users are always able to access the latest published version and get the most up to date developments, or instead use a specifically dated release, and maintain stability of their own resource. They can also test different versions and ensure no conflicts are created between versions before deciding to update. The ID policy (`http://obofoundry.org/id-policy.shtml` has been adopted throughout the OBO Foundry

[11] `http://code.google.com/p/oboformat/`
[12] `http://www.ontobee.org/`

6.2.2 Improving documentation by sharing metadata through the IAO

The IAO is an ontology of information entities, which aims at providing high-level blocks upon which specific resources can build. It describes classes such as *directive information entity*, which can for example be extended in a clinical-focused ontology by the *clinical guideline* subclass (see 8.3.3 for an example in AERO). As part of the IAO project, a distinct file defining common metadata properties[13] has been created. This file can be imported independently of the "core" IAO, and used by any developer. The IAO common metadata set contributes to the realization of the principle of documenting ontologies within the OBO Foundry.

Other efforts already exist to formalize metadata, such as the Simple Knowledge Organization System (SKOS) [162] and the Dublin Core (DC) Metadata element set [163]. However, considering the case of `dc:creator`, its definition reads "An entity primarily responsible for making the resource", where the resource is described by the class bearing this property. For example, in a book description, the `dc:creator` property value is set to the name of the author of the book, and does not capture the name of the author of the book description, which is what is intended with `iao:definition_editor`[14]. Similarly, the definition of `skos:definition` defines concepts, which is not suitable in this case [164].

In the IAO metadata set, common and expected annotation properties, such as *definition* and *editor preferred term* are documented, and allow tool developers to rely on them to build their user interface. Other properties such as *definition source* or *definition editor* were created to store any references used in developing the definition and who did create the term. This allows resource consumers to go back and check on the origin of the term and what its intended meaning is, and/or contact the relevant individual should they need more clarification about its usage. The importance of having human readable definitions was described in section 2.2.3: for example, the AERO relies on the PHAC glossary, and includes references to the appropriate source via the *definition source* annotation property.

Curators of the ontology can add *example of usage* and *editor note* to further clarify what the term denotes and what its intended usage is. Other slightly more complex properties have been designed to enable quality assessment of the terms. Namely, the *curation status specification* class provides a list of predefined instances (i.e., *'example to be eventually removed', 'metadata complete', 'organizational term', 'ready for release', 'metadata incomplete', 'uncurated', 'pending final vetting',*

[13] `http://purl.obolibrary.org/obo/iao/ontology-metadata.owl`

[14] `http://dublincore.org/documents/dcmes-xml/`, section 2.4

'to be replaced with external ontology term', 'requires discussion'[15]) that can be used on each class to mark its degree of "readiness" and stability. Similarly, the class *obsolescence reason specification* offers a list of predefined values that can be used on obsoleted terms to give more information as to why that term was deprecated and indicate (in conjunction with for example an editor note) what the term replacement is. Finally, an *OBO Foundry unique label* annotation property (`http://purl.obolibrary.org/obo/IAO_0000589`), was recently added in the ontology-metadata file to allow disambiguation between terms local to a resource when they are taken in the whole set of OBO Foundry ontologies. *OBO foundry unique labels* are automatically generated based on regular expressions provided by each ontology, when processed by the OBO package manager currently being written by the OBO Foundry custodians. Appendix E provides further description of the IAO annotation properties. The IAO metadata set is distributed as a file independently of the main IAO (which deals with representation of information entities), allowing resources to selectively import this ontology-metadata file.

6.2.3 Discussion

Despite the progress made on homogenizing some aspects of the OBO Foundry consortium guidelines, work remains to be done in several aspects. For example, sometimes terms need to be retired as ontologies evolve. The OBO Foundry doesn't currently formalize a standard deprecation policy, which leads to the problem of different policies within resources. As a general guideline, deprecated terms are not deleted from the ontology: (1) deleting terms contravenes the Cimino desiderata presented in section 2.3 and (2) removing a term that has been used in the past can be confusing for users. Some discrepancies exist between the practice of the GO [165] and other resources, such as OBI: in the GO [166], when terms are merged one term effectively disappears from the ontology file and its identifier is maintained as an *alt_id* annotation property on the term it is merged with. By contrast in the OBI, one term is deprecated, and its *obsolescence reason specification* is set to "term merged", with the addition of an editor note indicating the replacement term. As a consequence, tools such as MIREOT expect to find the URI of classes in their declaration (and not as a secondary ID). MIREOT scripts are therefore unable to retrieve the external information in the GO merging case, resulting in a loss of terms on the importing ontology side, such as

[15] `http://code.google.com/p/information-artifact-ontology/wiki/OntologyMetadata`

recently happened with some PATO terms [128][16]. A common deprecation policy, following the example of what has been done regarding the ID policy, would help formalize expected behavior, and guide tools developers. A review of the current reasons for obsolescence in the GO would be useful to perform to ensure adequacy between the instances defined by the IAO and the needs of the curators. Most proposed policies have been adopted fairly recently, and evaluation is very preliminary. Although the relative costs and benefits could be difficult to quantify, a number of use cases illustrate the advantage of relying on numerical identifiers. When choosing to use numerical IDs for terms, it is anticipated that tooling issues will hinder adoption of those standards - nobody wants to type in OBI_0001234 when doing a SPARQL query. However, it is believed that in the long term (i) tooling issues will be resolved (ii) using numerical IDs will be beneficial for maintenance of the resources and their necessary evolution. As illustration of these respective points, see for example the recent threads mentioning how (i) the Protege [126] team added a new menu "render by rdfs:label" to their interface [17] and (ii) issues faced by the developer of GoodRelations [167] to rename some classes.[18] Those policies also need to evolve with time and accommodate for example legacy resources. An update has been recently proposed [19] to enable the Protein Ontology (PRO) to reuse identifiers from existing databases, such as UniProt [168] in the interest of (1) making the connection to the original resource more explicit (2) not having to mint new identifiers for each existing identifier in the UniProt database. However, corollary to that update, is managing the dereferencing of those additional terms, which implies correspondingly updating the redirection rules to accommodate the new identifier format [20], without breaking existing support. A description of the current (November 2013) redirection rule is available at `http://code.google.com/p/obo-foundry-operations-committee/wiki/OBOPURLDomain`.

6.3 The MIREOT guideline

6.3.1 Introduction

The ability to share and reuse existing ontological resources is an important consideration when developing a new ontology. For example, when developing an ontology related to the biomedical

[16] `http://sourceforge.net/mailarchive/forum.php?thread_name=99D14FA3-9952-4C67-B892-41A8499A43C8%40gmail.com&forum_name=obi-devel`

[17] `https://mailman.stanford.edu/pipermail/p4-feedback/2011-May/003889.html`

[18] `http://lists.w3.org/Archives/Public/public-lod/2011Apr/0278.html`

[19] `http://code.google.com/p/obo-foundry-operations-committee/issues/detail?id=81`

[20] `http://code.google.com/p/obo-foundry-operations-committee/issues/detail?id=88`

domain, it may be useful to include terms from the GO [165] to represent biological processes or from the PATO [128] to represent properties of entities. Ontologies such as GO and PATO are built collaboratively by communities of experts and are the products of substantial effort. Recapitulating this work instead of reusing it represents a duplication of development effort and results in multiple ontologies covering the same domain. It could also result in projects having different unique identifiers to denote the same entity, which would require post-hoc, potentially error-prone, identifier mapping systems to enable data integration.

While it appears that building upon existing ontologies is the best way to proceed, developers are faced with a number of practical challenges when trying to do so. The easiest way to integrate an existing ontology is to rely on the *owl:imports* [160] mechanism, which imports the external resource as a whole. However, current limitations in tools and reasoners can sometimes make this impractical. Popular OWL tools (e.g., Protégé [126] and Pellet [60]) can neither load nor reason over very large ontologies such as the NCBI Taxonomy Database [169] or the Foundational Model of Anatomy [170]. Furthermore, external ontologies may have been constructed using design principles which do not align with the principles of the ontology requiring their import. In this instance, wholly importing such ontologies could lead to inconsistencies or unintended inferences [171]. Other import options are possible, for instance using software that extracts a *module* [172] of the external ontology.

A module can be seen as a subset of an external ontology that, when imported by another ontology, allows the same inferences to be drawn with respect to the classes of interest as if the whole ontology had been imported, and answer queries without losing any reasoning power. However, if an extracted module is to be useful, the external ontology needs to be structured in a way that is compatible with the importing ontology (e.g., using the same upper ontology and relationship types), and the logical axioms need to accurately represent existing knowledge, which is not always the case at the current stage of development of some resources. For example, during the development of the OBI [83], importing the root class of the Common Anatomy Reference Ontology (CARO) [173] was not desired as its definition intersected multiple classes in OBI, making it difficult to determine how the two ontologies aligned. Specifically, the root of CARO, *anatomical entity*, encompasses *material* and *immaterial material entity*, which belong to different hierarchies in OBI. In addition, although software that extracts modules are available, most are in early stages of development.

Several modularization tools [60, 174, 175, 176] were tried. All of them discarded annotations, resulting in modules containing only the class declarations and no annotation properties, such

as labels or definitions. There were also software crashes on large ontologies (the size of the ontologies capable of being loaded varying with the tool; for example the Chemical Entities of Biological Interest (ChEBI) Ontology [155] can be loaded with SWOOP but not with Protégé 3.4). One tool [176] had undocumented assumptions about the form of URIs used as class names and therefore extracted empty modules. The other tools described were able to extract modules by automatically determining their size. This resulted in either a single term or a large number of terms being extracted, depending on the provided arguments, as the tools attempt to approximate a module without discarding potentially useful information. These large modules undermine the goal of having imports of a manageable size. In conclusion, the current ontology modularization tool set is in the early stages of development and, though promising, does not address current needs.

To address these issues a set of guidelines for importing terms from multiple resources, avoiding the overhead of importing the complete ontology from which the terms derive was developed. In collaboration with Alan Ruttenberg, I created the MIREOT guidelines to aid the development of OBI. OBI uses the BFO [79] as an upper-level ontology and has been submitted for inclusion in the OBO Foundry [55]. MIREOT enables reuse, where appropriate, of existing ontology resources, therefore avoiding duplication of effort and ensuring orthogonality (i.e., non overlapping scope), and contributing to the realization of a fundamental principle of the OBO Foundry. MIREOT is a guideline independent of any design principle, and provides a mechanism by which external ontology terms can be selectively imported, even if they do not use a particular upper ontology or OWL DL [109].

6.3.2 Policy

In deciding upon a minimum unit of import, the first step was to consider the practice of other ontology efforts. For example, in the GO, the intended denotation of classes remains stable such that even when the ontology is repaired or reorganized, the effects of such changes do not affect the intended meaning of individual terms. Rather, the changes are towards more carefully expressing the logical relations between them. When a term's meaning changes, the term is deprecated [166]. Therefore a term can be considered as stable, in isolation from the rest of the ontology, and terms (i.e., individual classes in isolation from the ontology) can be used as basic unit of import. The current implementation of MIREOT has been limited to the import of terms from other ontologies that aspire to be a part of the OBO Foundry, and so adhere to a similar deprecation policy. The minimum amount of information needed to reference an external term is its URI (i.e., the identifier

for this term) and its source ontology URI (i.e., where the term comes from). Generally, these items remain stable and can be used to unambiguously reference the external term. The minimum amount of information needed to then integrate this class in the importing ontology is its desired position in the hierarchy, specifically the URI of its direct superclass (i.e., under which class the term is to be asserted).

Taken together, the following minimal set is enough to consistently reference an external term:

1. **Source ontology URI** The logical URI of the ontology containing the external term to be imported.
2. **Source term URI** The logical URI of the specific term to import.
3. **Target direct superclass URI** The logical URI of the direct asserted superclass in the importing ontology.

While physical URIs may evolve over time, logical URIs are stable and can be used to unambiguously refer to the same term. To ease development of the importing ontology, it is also recommended, although not required, that additional information about the external class be added, such as its label and textual definition, or any other kind of information that may be deemed useful by the ontology developers. This additional information, when appropriate, is mapped into the importing ontology's annotation properties. As it is prone to modification by the source ontology developers (*e.g.*, when updating a definition), it is stored in a separate file that can be removed and rebuilt on a regular basis, allowing for regular updates within the importing target ontology.

6.3.3 Implementation

I performed an implementation of the MIREOT guidelines in the context of the OBI project (Figure 6.1), and can be decomposed into a two-step process:

1. Gather the minimum information for the external class.
2. Use this minimum information to fetch additional elements, like labels and definitions.

Once the external term is identified for import, the first step is to gather the corresponding minimum information set. This set is stored in a file called *external.owl* (all scripts and files are available under the OBI Subversion Repository [177]). In the current implementation, a Perl script, *add-to-external.pl*, can be used to append the minimum information set for a given external term

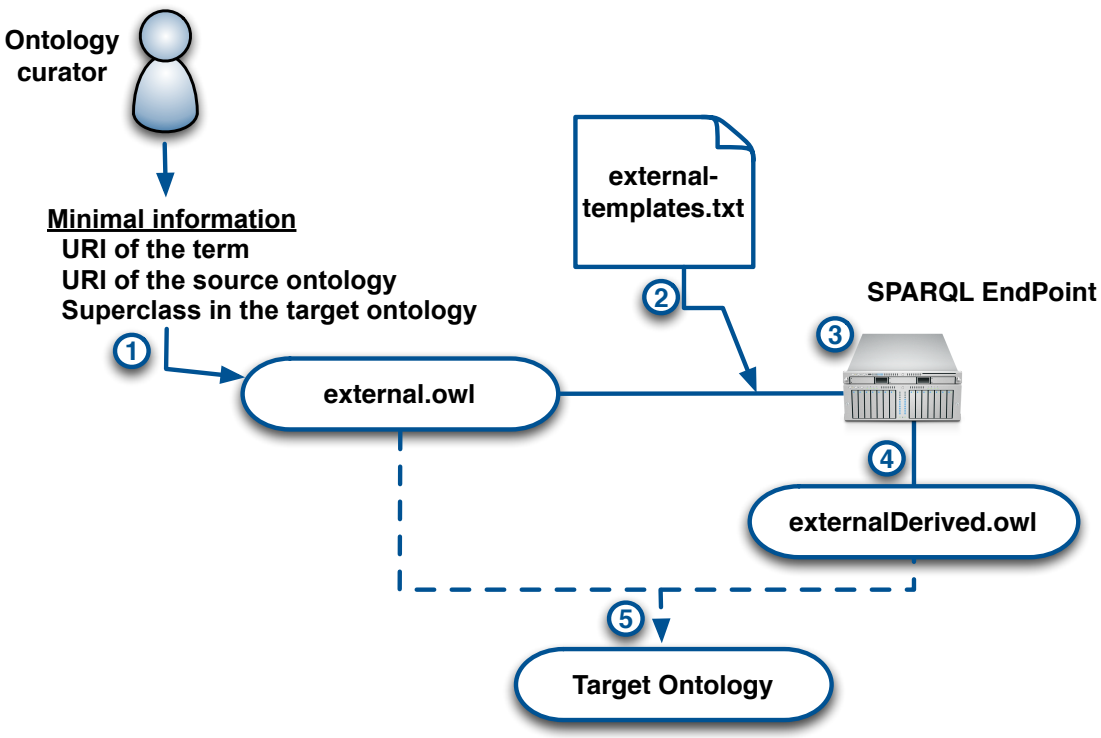


Figure 6.1: Diagram of the MIREOT mechanism as implemented by OBI. 1. The ontology editor gathers the minimal information for the class to import and adds it into the external.owl file 2. A script parses the external.owl file, and for each class selects the appropriate SPARQL CONSTRUCT template. 3. The SPARQL query is executed against a SPARQL endpoint (e.g., Neurocommons) 4. The results of the SPARQL queries are combined into the externalDerived.owl file 5. The target ontology imports the external.owl and externalDerived.owl files.

to the *external.owl* file. The script takes as arguments the identifier of the external class to be imported and its parent class in the target hierarchy. In addition, a mapping between the prefix used in the identifier and the external source ontology URI is built into the script. For example, when requesting the term *CL:0000767* (see below), the script maps the *CL:* prefix to its source ontology URI `http://purl.org/obo/owl/CL`. Curators therefore need only specify the ID of the external class to import (rather than the full URI) and the ID of the class it should be imported under. Upon addition of an external class, a visual check can be performed as the Perl script returns to standard output the OWL excerpt added to the file.

In the current implementation, the additional information can be obtained programmatically via SPARQL [94] CONSTRUCT queries (Figure 6.2). While access to a SPARQL endpoint is not

compulsory to use the MIREOT mechanism, it provides easy access, using standard protocols, to the information needed. These queries [178] specify, for each source ontology, which extra elements about the term is to be extracted, such as the definition and preferred label, and how to map these into the corresponding OBI annotation properties.

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix obi: <http://purl.obofoundry.org/obo/>
prefix obo: <http://www.geneontology.org/formats/oboInOwl#>
prefix iao: <http://purl.obolibrary.org/obo/>

construct
{
  _ID_GOES_HERE_ rdf:type owl:Class.
  _ID_GOES_HERE_ iao:IAO_0000111 ?label.
  _ID_GOES_HERE_ rdfs:label ?label.
  _ID_GOES_HERE_ iao:IAO_0000115 ?definition.
}
where
{
  {   _ID_GOES_HERE_ rdfs:label ?label. }
  UNION
  {   _ID_GOES_HERE_ obo:hasDefinition ?blank.
      ?blank rdfs:label ?definition}
}
```

Figure 6.2: Template SPARQL query. For convenience, alias:preferredTerm and alias:definition are used to reference annotations properties IAO_0000111 and IAO_0000115 [125] respectively. The _ID_GOES_HERE_ pattern will be replaced by the script when building the CONSTRUCT query.

For example, in the current OWL rendering of OBO files, definitions are individuals and the rdfs:label of those individuals records the text of the definitions. Within the OBI implementation of the MIREOT guidelines, the value of the rdfs:label of the oboInOwl:Definition will be set to the value of `http://purl.obolibrary.org/IAO_0000115` (i.e., iao:definition). Only annotation properties which map directly to the target ontology's own metadata are copied; new properties, if not specified in the source ontology, are not created.

Finally, a script, *create-external-derived.lisp*, iterates through the minimum information stored in *external.owl*. Depending on the source ontology URI of each of the imported terms, it then selects the correct SPARQL template and substitutes the relevant ID. The queries are then executed against the Neurocommons OBO SPARQL endpoint [57, 179]. This supplementary information is stored in a second file, *externalDerived.owl*. This file can be removed on an ad-hoc basis (e.g., before releasing new versions of the importing ontology) so that it can then be rebuilt via script based on *external.owl* in order to refresh the additional information (e.g., label). The two files, *external.owl* and *externalDerived.owl*, are then imported by the target ontology, providing the necessary information to the editors while at the same time keeping it independent from the importing ontology's proprietary classes. This introduces an additional level of modularity, separating the domain ontology of interest from the external ontologies.

In the following sections I present three different cases of application of the MIREOT guidelines, implemented during the OBI development.

Use Case One - Basophil and Cell classes

The OBI *cell* class was replaced with that from the Cell Type (CL) ontology [180]. CL is part of the OBO Foundry effort, and the *cell* class as defined by this resource should be reused, instead of creating another class denoting the same entity. This class can subsequently be chosen as the parent of another imported term as needed. For example the following invocation of the *add-to-external.pl* script:

```
perl add-to-external.pl CL:0000767 CL:0000000
```

will add the class *basophil* (CL:0000767) as subclass of the class *cell* (CL:0000000), and set the source ontology URI as `http://purl.org/obo/owl/CL`. Once imported, the *basophil* and *cell* classes can be used like any other OBI class. For example, the material entity *CD3+ T cell culture* is defined as:

```
Class: CD3+ T cell culture
    SubClassOf:' cell culture'
                          and 'has grain' some (cell
                          and (has_part some 'CD3 subunit with immunoglobulin domain'))
```

Use Case Two - taxonomic information

The *cell* use-case highlights what is likely to be the most common import scenario (i.e., a simple import of one external term, making it available for direct use in the target ontology). However, in some cases, more than that single external term may be required, and to account for this MIREOT has been devised to be flexible.

Consider the scenario in which there are two experiments, one in human and one in mouse. The files are annotated with the classes *human* and *mouse* from the ontology, which are in turn mapped from the NCBI taxonomy database. Somebody could want to query for all experiments in mammals, without specifying the exact species. In this case, one needs to know that *human* and *mouse* are subclasses (even indirect) of mammals in the NCBI taxonomy. The root term of the NCBI taxonomy database is an example of a term OBI didn't want to include, as it encompasses *viroids*, *unclassified sequences* and *others sequences*, which were not considered useful when defining organisms. Therefore, when mapping towards an NCBI term, it was decided to also retrieve all its superclasses up to the *Archaea*, *Bacteria*, *Eukaryota* and *Viruses* levels of the NCBI taxonomy database (Figure 6.3). When the *create-external-derived.lisp* script parses the *external.owl* file and encounters an NCBI taxonomy ID, it will invoke a specific SPARQL query (Figure 6.3). As per the mechanism described above, the minimum information about the imported external class (e.g., *Mus musculus*) is defined in *external.owl*, whereas the additional rank information (e.g., genus, kingdom, phylum i.e., its superclasses) is stored in *externalDerived.owl*. On the same model, any information that the importing ontology editors would require could be added in the *externalDerived.owl* file: the only requirement is to write the corresponding SPARQL query.

Use Case Three - Unit instances

Finally, the most recent use case addresses the needs for OBI to represent units of measurement. The Unit Ontology (UO) [156] tackles this effort, and currently encompasses more than 2000 classes.

However, the representation of units as classes doesn't comply with the design pattern chosen by OBI and the IAO, which take the stance that in the absence of a satisfactory unit representation theory, things that are understood, i.e., unit labels, should be represented. Therefore the UO classes corresponding to specific units (such as *gram* or *meter*) were imported as instances of the IAO class *measurement unit label*. Figure 6.4 shows the result of this addition into the OBI hierarchy.

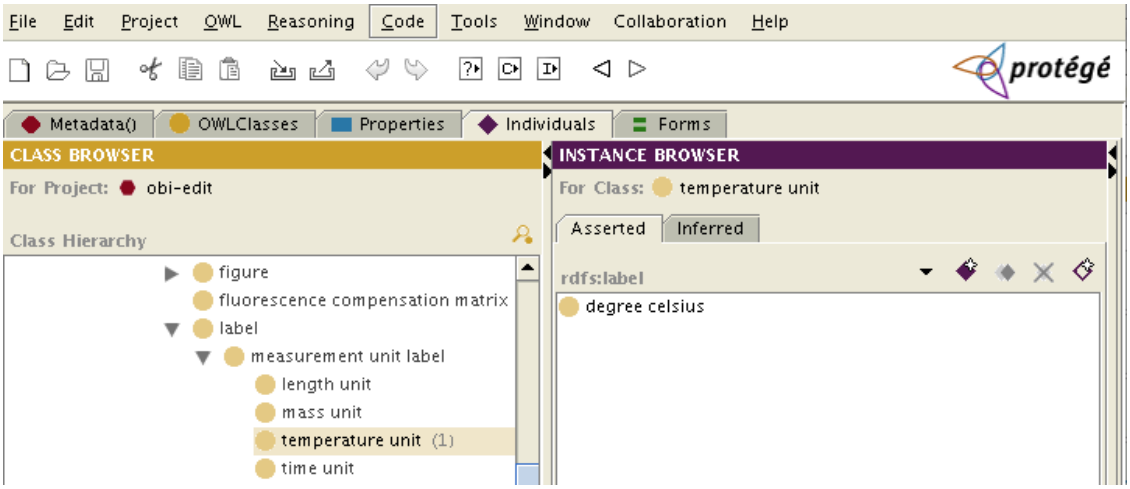


Figure 6.4: Screenshot of the Protege editor showing the class *temperature unit* and its instance *degree celsius*, as imported using the MIREOT mechanism from the UO ontology.

Work is in progress with the developers of the UO to reach agreement on the best way to represent measurement units in a consistent manner, and it is expected the different resources will align as part of the OBO Foundry collaborative effort.

6.3.4 Discussion

The MIREOT mechanism offers a lightweight mechanism for importing specific classes from external ontologies. The approach is decoupled from the importing ontology, allowing a computational update mechanism which does not interfere with the primary ontology under development. MIREOT is currently implemented and used by several ontology efforts, including OBI, the IAO, the VO [119], the IDO [131] and the Influenza Ontology (InfluenzO) [136]. In the context of OBI, 472 terms are currently explicitly imported, which in turn leads to actual integration of 1447 classes (due to the automatic retrieval of parents when using the NCBI taxonomy).

With broader use of the MIREOT mechanism by OBI and other resources, several minor issues arose. The first issue is a case of cyclic imports between resources: for example, IAO developers

required import of the term *investigation*, which class already exists in OBI. However, OBI imports IAO, and therefore re-imports, *via IAO*, its own *investigation* class. This is not problematic in general, as duplication of triples in OWL files is of no consequence. However when OBI curators decided to update the definition of the *investigation* class, the information natively in OBI and that imported from IAO became out-of-sync: two different definitions were displayed to the curators. Moreover one of them could not be edited as it is outside the remit of OBI to edit IAO definitions. One solution to this problem would be to update the IAO import - but this requires a release of OBI with the updated *investigation* definition, its upload on Neurocommons, and for the IAO developers to update their information and produce a new release of IAO. At best, this implies a delay of a few days, more realistically of a few weeks until the information in both files is again synchronized. Such a solution also has consequences; when updating the information from the SPARQL endpoint, a specification of which RDF graph [181] the term *originally* belonged to is required. Taking again the example of the *investigation* class, when querying based on its URI without specifying the RDF graph, the OBI class, but also the one distributed by IAO, will be returned. This is not the desired behavior; in this example, the IAO annotation property values are now out of date compared to the original and authoritative OBI file. A better solution would be for tools to recognize and prioritize the origin of a class based on its URI. Ontology editing tools would then display only the information originating from the target ontology when editing the target ontology file. This issue remains to be addressed.

Additionally, when updating imported information, the SPARQL endpoint where the information resides must be up-to-date. The implementation currently relies on the OBO Foundry resources at the Neurocommons OBO distribution. This is updated nightly with the latest information from the OBO server, and can therefore be reasonably relied upon for accessing current resources. The timeliness of the information may not always be known if extending the mechanism to another SPARQL endpoint, or other sets of ontological resources. The MIREOT standard presents an approach to importing classes from external ontologies that removes the overhead of full ontology imports whilst maintaining a decoupled but usable reference to the external classes. There is a clear trade-off that MIREOT offers between practicality and full, axiomatic completeness. Being a lightweight import mechanism, only the desired parts of an external ontology are imported, at the risk that inferences drawn may be incomplete or incorrect; correct inference using the external classes is only guaranteed if the full ontology, or a module, is imported. It does however present the important advantage of overcoming the obstacle presented by ontologies which are not fully interop-

erable at present. Since only partial, reasoner-supported consistency checking is undertaken, extra care is taken when assertions about an imported term are made. In adding axioms, such as the subclass axiom when importing the external term, the aim for the ontology editor is to only assert true statements, which do not contradict or alter the meaning of the term in its source ontology. With the more stable OBO ontologies, the denotation of the term, as explained in the definition or documentation, is clearer and more correct than the axiomatization, the former being easier and quicker to formulate. It is anticipated that some of the statements added by the importing ontology may migrate to the source ontologies at some point in the future; a fruit of the collaborative nature of OBO Foundry ontology development.

When deciding to import an external term the textual definition is reviewed and, if required (e.g., if the definition is ambiguous), discussion with the original editor is undertaken. An important aspect of the MIREOT mechanism is maintaining the term's meaning, and ensuring that if it changes the term gets deprecated, and therefore it is recommended to use resources adhering to a deprecation policy. As imports are done from OBO Foundry candidate ontologies there is a community process for monitoring change, a shared understanding of the basics of the domain, and the intention to eventually share the same upper-level ontology. Therefore, it is expected that terms will be deprecated if there is a significant change in meaning, and the MIREOT mechanism is flexible enough to adjust and update import of terms as the other ontologies start enhancing their logical definitions.

Finally, the original implementation of the MIREOT guidelines relied on command-line scripts and specific libraries, making it difficult for curators with no programmatic skills to use. Subsequently, a web service, OntoFox [182], described in the following section, has been developed to facilitate the process.

6.4 OntoFox

6.4.1 Introduction

MIREOT is being used in an increasing number of ontology projects, for example, OBI, VO, the InfluenzO - `http://sourceforge.net/projects/influenzo/`, Neural ElectroMagnetic Ontologies (NEMO) - `http://nemo.nic.uoregon.edu/wiki/NEMO`, ontologies developed in the Neuroscience Information Framework (NIF) - `https://confluence.crbs.ucsd.edu/display/NIF/`, and as part of the eagle-i project (`https://www.eagle-i.org/home/`). While editing tools commonly provide

means to reference an external term by directly setting its URI, one must also manually enter auxiliary information necessary for practical editing, such as the label and definition, and update such information if the source ontology changes. Manual entry is time consuming and duplicates already existing information. Also, such terms would be hard to distinguish from those in the current resources, making their update a tedious process. In addition, it is often desirable to import additional related terms. For example, when the VO imports a species term, the inclusion of some of its superclasses allows for queries at different taxonomic ranks (e.g., kingdom, phylum, and species). To address these issues, an initial implementation based on MIREOT was created to facilitate managing the tedious aspects of this process automatically.

Alternatives such as computing modules [183] were investigated. Structural approaches use the syntax of the axioms of ontologies and mostly only consider the induced is-a hierarchy [176, 184]. Logic-based approaches take into account the consequences of ontologies and require that this extracted module captures the meaning of the imported terms used, i.e., includes all axioms relevant to the meaning of these terms. However, Grau et al. [172] proved that it is undecidable, even for description logics simpler than OWL-DL, to determine whether a subset of an ontology is a minimal logic-based module. These approaches are relatively new, experience using them is limited, and experience with current Web-based implementations has found them to be unreliable. Moreover the methods do not provide ways to avoid import of certain terms or axioms that might not be considered desirable, or have other issues that prevent their easy use. Nonetheless the syntactic locality approach these methods use is applicable to single-term import and so is compatible with the MIREOT approach.

In section 6.3, an implementation of the MIREOT mechanism that demonstrates the feasibility of the approach is described. It is, however, command line-based and requires the specification of terms either by command-line scripts or construction of an ontology document. Specification of which ancillary information should be incorporated is by writing SPARQL queries, restricting its adoption by less technically able users. To facilitate application of the MIREOT guideline by the wider ontology community a more user-friendly system facilitating the import and update of external terms into a target ontology is desired. In addition, while MIREOT provides a practical yet simple approach to specifying external ontology terms, the OBI implementation does not provide the ability to consider restrictions on imported terms that a user may desire to import. To preserve the meaning of the imported terms, ontology developers might like to use ontology module extraction, e.g., extraction of the target class and its transitively related (via restriction) closure [176]. Ontology

developers may also want the flexibility of including no superclasses, only one direct superclass, all superclasses to the top class, or a subset of all superclasses for a term, in order to provide additional relevant domain terms for their users.

To address these needs for ontology reuse, OntoFox (`http://ontofox.hegroup.org/`), a web-based application implementing the MIREOT and related ontology term extraction strategies was developed. OntoFox facilitates ontology development by automatically fetching properties, annotations, and related terms from external ontology terms and saving the results as OWL serialized as RDF/XML [181] suitable for use with the OWL import directive. OntoFox provides a web-based package of solutions for ontology developers to extract, for subsequent import, different sets of ontology terms by following and expanding the initial MIREOT implementation and by developing related ontology term extraction methods based on SPARQL [94]. The following sections describe the general OntoFox web system, how users can choose which properties and related terms should be imported, and demonstrate how OntoFox is used in the VO development.

6.4.2 Methods

OntoFox system architecture

OntoFox uses a simple text format and web forms for data input in a user-friendly implementation, and is designed to not require any programming skills. OntoFox is implemented using a three-tier system architecture. At the front-end, data can be submitted using either web forms or by uploading a plain text input file. The input data are then processed using PHP and Java, and SPARQL (middle-tier, application server) queries are then executed against an RDF triple store (back-end, database server), currently the Neurocommons SPARQL endpoint [67]. The web server then processes the result of each SPARQL query sent by the back-end server; as a result an RDF/XML file is created and offered for download to the user.

As OntoFox is a web-based system, it is accessible everywhere through the Internet without need for additional software installation. The techniques used in the OntoFox web application were chosen for maximum compatibility by using established W3C standards, specifically, OWL as a web ontology language, RDF/XML as its serialization, and SPARQL for queries.

OntoFox three-tier structure implementation

1. *OntoFox web interface* The OntoFox web interface is designed based on iterative testing, thus far informal usability testing and feedback from initial users, following a spiral software development model [27]. It accepts the input from the user, via either web forms or uploading of a local text file, and presents the output data after query processing. Finding and entering the URIs for desired terms can be tedious. To speed up the term specification process, an ontology term suggestion feature based on auto-completion of the string of text entered by users after selecting the desired source ontology was implemented. The OntoFox server offers a list of potential matches, and upon selection, the associated term ID will show up in an input box next to the label. Additionally, the "Detail" hyperlink next to the term ID provides easy access to an interactive ontology browser allowing visual confirmation of the term definition and its position in the hierarchical ontology tree structure. Lastly, as shown in Figure 6.5, the user can click "Add" next to "Detail" to insert the full URL of the selected term into the input text box on the web interface.

```
# give names to the top taxa
alias:bacteria=tax:_2
alias:eukaryota=tax:_2759
alias:archaea=tax:_2157
alias:viruses=tax:_10239
alias:cellularOrganism=tax:_131567

prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix obi: <http://purl.obofoundry.org/obo/>
prefix tax: <http://purl.org/obo/owl/NCBITaxon#NCBITaxon>
prefix iao: <http://purl.obolibrary.org/obo/>

construct
 { ?super rdf:type owl:Class.
   ?super rdfs:subClassOf ?parent.
   ?super iao:IAO_0000111 ?label.
   ?super rdfs:label ?label.
   ?super alias:importedFrom <http://purl.org/obo/owl/NCBITaxon>
 }
where
{
 { # We harvest the transitive superclass annotations
    _ID_GOES_HERE_ rdfs:subClassOf ?super.
    graph <http://purl.org/science/graph/obo/NCBITaxon>
     { ?super rdfs:subClassOf ?parent.
       ?super rdfs:label ?label.
     }
  }
UNION

 { graph <http://purl.org/science/graph/obo/NCBITaxon>
   { ?super rdfs:subClassOf ?parent.
     ?super rdfs:label ?label.
     FILTER (?super=_ID_GOES_HERE_)
   }
  }

FILTER (!((?super=alias:bacteria) || (?super=alias:eukaryota) || (?super=alias:viruses)
 || (?super=alias:archaea)
 || (?super = alias:cellularOrganism) || (?parent = alias:cellularOrganism)))
}
```

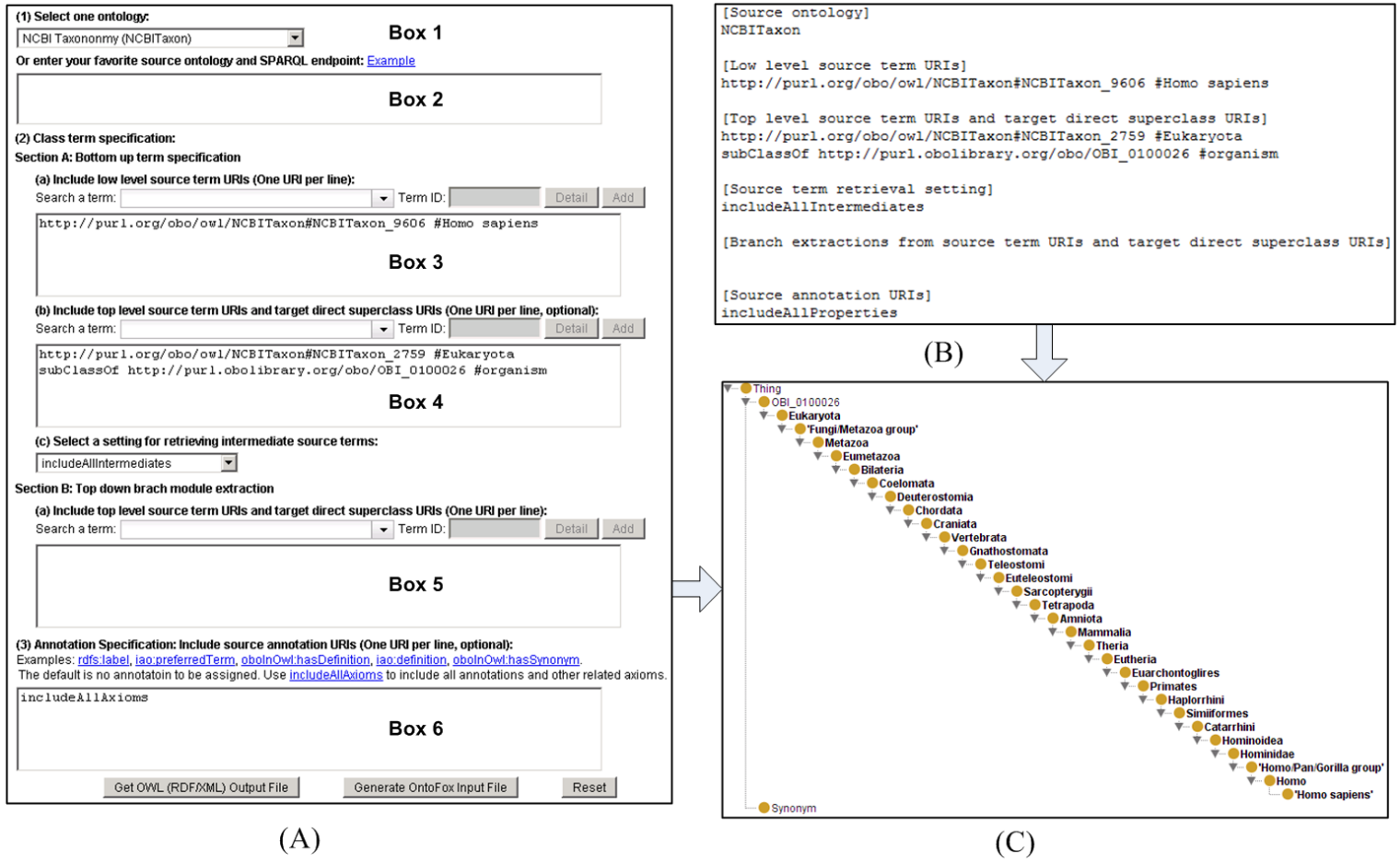


Figure 6.5: OntoFox retrieval of the term 'homo sapiens' from the NCBI Taxonomy Ontology (NCBITaxon). Input data can be entered via web-based forms (A) or text file upload (B). The output OWL file [Additional file 1] can be visualized using Protégé (C). All terms from 'homo sapiens' up to Eukaryota are retrieved. Synonyms used to annotate each term are also included.

2. *Data processing by the web application* The OntoFox application server runs on a Dell PowerEdge 2580 server running the Red Hat Linux operating system (Red Hat Enterprise Linux 5 server). PHP and Java are used as programming languages in the web application server. General web-based programming and query submission are written using PHP. The OWL API [150], a Java API for manipulating OWL files, is used in OntoFox to read, process, and rewrite OWL files and save the final results as one OWL file after merging individual query results.

3. *Data storage and access* The OntoFox internal RDF database server runs on a separate Dell PowerEdge 2580 server. The database server is powered by the OpenLink Virtuoso database engine [103]. While VO is loaded within this Virtuoso server, OntoFox also uses RDF data stored in other web accessible servers, for example, the Neurocommons knowledge management platform [57]. Fifteen biomedical ontologies generally used within the OBO community are available for users to select as source ontologies within OntoFox (Table 6.1). These ontologies, initially chosen to support VO development, were selected based on their specificity, community support, and maturity. They all adhere to a strict deprecation policy, ensuring that the meaning of each term remains stable until the term is deprecated. Though nothing bars serving more resources, these 15 ontologies were all that were required to cover all information needed for import via MIREOT during the VO development. Users can choose to provide another source ontology URI and corresponding SPARQL endpoint, allowing retrieval of terms outside of the OntoFox source ontology repository resources; however this is done at their own risk as term stability is not guaranteed.

Evaluation of OntoFox SPARQL retrieval of related terms

To compare the performance of the OntoFox SPARQL related term retrieval approach with the OWLAPI modularization, three sets of signature data were used. I performed the OWLAPI modularization, while Zuoshuang Xiang ran the queries on OntoFox. The first two sets of signature data include either one term (e.g., the OBI term 'antigen') or a list of OBI terms that were imported to VO. The third set of signature terms for modularization includes all terms in the NIF Lexicon ontology (nif.owl; `http://ontology.neuinfo.org/NIF/nif.owl`). The nif.owl file uses approximately 30 external files. The OntoFox method and the OWLAPI modularization method were separately performed and compared. For the OWLAPI modularization, the OWLAPI Syn-

tacticLocalityModuleExtractor with STAR module type was used.

6.4.3 Results

MIREOT implementation

As described in Section 6.3, the MIREOT guideline suggests the following minimal set: (1) source term URI, (2) target direct superclass URI, and (3) source ontology URI. These are the first parameters taken as input by OntoFox:

1. *Source ontology URI.* Box 1 of the OntoFox web input system includes a list of the 15 ontologies a user can select as source ontology (Figure 6.5). Alternatively, a user can request an unlisted source ontology in Box 2, in which case the URL of a SPARQL endpoint where this new source ontology can be accessed must be provided. For each external ontology term, OntoFox adds an importedFrom annotation property (`http://purl.obolibrary.org/obo/IAO_0000412`), which indicates the URI of the source ontology.

2. *Low level source term URI.* This parameter is equivalent to the source term URI in the MIREOT guideline. Box 3 allows users to input one or multiple source term URIs, entering one URI per line. For example: `http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9606` #Homo sapiens

3. *Target direct superclass URI.* This is the URI of the direct superclass of the top-level source term chosen above (i.e., where to position the newly imported term(s) in the target ontology). This parameter is entered alongside the top-level source term URI in Box 4 using the directive "subClassOf" (see more detail below).

These three data items together unambiguously define a single term from the source ontology and where to position (i.e., what class is it a subclass of) it in the target ontology.

Annotation properties management

OntoFox provides several settings/directives allowing users to select which annotation properties to retrieve, and more importantly, under which format those should be returned.

1. Source term annotation URIs: By default (i.e., if no annotation URI is specified), OntoFox will not fetch any of the annotation properties of the selected term. A user can choose to

retrieve specific annotation properties by specifying their URIs, or use the OntoFox command 'includeAllAxioms' to fetch all annotations properties associated with source ontology terms. This parameter is entered in Box 6 in the web input format (Figure 6.5).

2. "copyTo": This directive is used to map an ontology term annotation to a new annotation property created in the target ontology, resulting in a duplication of the annotation property value in the output file. It is used at the beginning of a line, followed by an annotation URI used in target ontology. For example, the "copyTo" command is used in Figure 6.6:

   `http://www.w3.org/2000/01/rdf-schema#label` copyTo `http://purl.obolibrary.org/obo/IAO_0000111` #preferred term

   This duplicates the value of the rdfs:label property into the "preferred_term" annotation (IAO_0000111 from the IAO [125]), and both annotations are included in the output file. This directive can be used in the web form (Box 6) or in the OntoFox input text file.

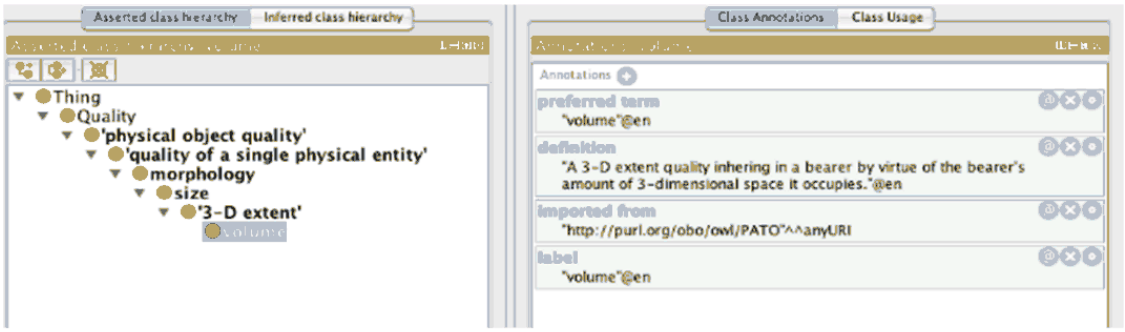
3. "mapTo": This directive allows mapping of an ontology term annotation: it will replace an existing annotation property in the target ontology with the value of another annotation property from the source ontology. It is used at the beginning of a line, followed by an annotation URI from the target ontology. For example, Figure 6.6 contains an example of using the "mapTo" directive:

   `http://www.geneontology.org/formats/oboInOwl#hasDefinition` mapTo `http://purl.obolibrary.org/obo/IAO_0000115` #definition annotation property

As ontologies don't always use a common set of annotation properties, this feature provides an easy way to integrate information from a source ontology into a target ontology while retaining a consistent, metadata style. For example, the OBO2OWL script (`http://www.berkeleybop.org/obo-conv.cgi`), used to automatically generate OWL version of OBO ontologies within the OBO Foundry, uses the property "hasDefinition" to relate a term to an instance whose rdfs:label is that term's definition. However VO uses the IAO metadata scheme (`http://code.google.com/p/information-artifact-ontology/wiki/OntologyMetadata`), and directly relates the term to its definition via the `http://purl.obolibrary.org/obo/IAO_0000115`, definition annotation property. The mapTo directive instructs OntoFox to map the definition used in the source to the value of the VO annotation property for definition. This mapping directive is used in Box 6 of the web form input method or in OntoFox input text format.

```
[Source ontology]
PATO

[Low level source term URIs]
http://purl.org/obo/owl/PATO#PATO_0000918 #volume

[Top level source term URIs and target direct superclass URIs]
http://purl.org/obo/owl/PATO#PATO_0001241 #quality of continuant
subClassOf http://www.ifomis.org/bfo/1.1/snap#Quality

[Source term retrieval setting]
includeAllIntermediates

[Source annotation URIs]
http://www.w3.org/2000/01/rdf-schema#label
copyTo http://purl.obolibrary.org/obo/IAO_0000111
http://www.geneontology.org/formats/oboInOwl#hasDefinition
mapTo http://purl.obolibrary.org/obo/IAO_0000115
```

(A)



(B)

Figure 6.6: OntoFox retrieval of PATO term 'volume' and its annotations. (A) OntoFox input data; (B) Protégé display of OntoFox output data. All terms from 'volume' up to 'quality of continuant' in PATO have been imported and positioned under the BFO term Quality. The desired annotation properties (IAO_0000111 'preferred term' and IAO_0000115 'definition') have been specified using OntoFox directives 'copyTo' and 'mapTo'.

Managing incorporation of related terms

OntoFox provides a number of mechanisms for selecting related terms for import, all based on structural approaches and that have been used within VO development. Methods are provided for selective retrieval of parent terms, transitive retrieval of restrictions inspired by structural-based modularization techniques, and the extraction of a subtree rooted at a given term. In this section these mechanisms are detailed. The setting "Top level source term URI", is designed to work in conjunction with another term specification when retrieving parent terms between lower and upper level source terms. A typical use is when importing some or all of the superclasses of a species term

to allow for queries at different taxonomic ranks (e.g., kingdom, phylum, and species). For example, in between 'homo sapiens' and Eukaryota (the chosen top-level term) in the NCBI Taxonomy, there are 27 intermediate terms (cf Figure 6.5). It would be very tedious to find, copy and then paste all those 29 terms into the new ontology. By specifying 'homo sapiens' as the low level source term, Eukaryota as the upper level source term, and the setting "includeAllIntermediates", the 27 intermediate terms are automatically retrieved by OntoFox (Figure 6.5). In addition to this retrieval of all parent terms, OntoFox uses an algorithm to compute and retrieve intermediate source terms that are the closest ancestors of more than one low-level source terms, and to remove intermediate terms that have only one parent term and one child term (Figure 6.7), leaving only terms that present alternatives for query. This setting, "includeComputedIntermediates", provides an option to reduce the number of extracted ontology terms by getting less intermediate ontology terms than that with the setting "includeAllIntermediates" (Figure 6.5), while still fulfilling many users' requirement.

```
input: Ontology input Oi
output: Ontology output Oo
hasMore = true
while (hasMore)
   hasMore=false
   for each classT in Oi
      if ((number of superclass of classT) = 1 && (number of subclass of classT) = 1)
         classTsup = (superclass of classT)
         classTsub = (subclass of classT)
         add classTsub as a subclass of classTsup
         delete classT
         apply changes to Oi
         hasMore=true
return Oi as Oo
```

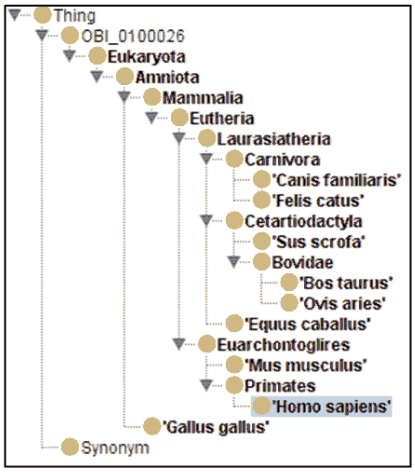
Figure 6.7: OntoFox algorithm for extracting computed intermediate classes. It removes any intermediate classes that have only one parent class and only child class. Only intermediate terms with at least two children classes are kept.

Figure 6.8 demonstrates the usage of this setting. 11 commonly used animal species are included as the low-level source terms. Using the setting "includeAllIntermediates", 70 intermediate terms will be included. However, only six intermediate terms are included after the "includeComputed-

Intermediates" setting is applied (Figure 6.8).

```
[Source ontology]
NCBITaxon

[Low level source term URIs]
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9606 #Homo sapiens (human)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_10090 #Mus musculus (house mouse)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9031 #Gallus gallus (chicken)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9685 #Felis catus (domestic cat)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9913 #Bos taurus (cattle)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9940 #Ovis aries (sheep)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9823 #Sus scrofa (pig)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9796 #Equus caballus (horse)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9615 #Canis familiaris (dog)
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_9443 #Primates
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_40674 #Mammalia (mammals)

[Top level source term URIs and target direct superclass URIs]
http://purl.org/obo/owl/NCBITaxon#NCBITaxon_2759 #Eukaryota
subClassOf http://purl.obolibrary.org/obo/OBI_0100026 #organism

[Source term retrieval setting]
includeComputedIntermediates

[Source annotation URIs]
http://www.w3.org/2000/01/rdf-schema#label
http://www.geneontology.org/formats/oboInOwl#hasSynonym
http://www.geneontology.org/formats/oboInOwl#hasExactSynonym
```

(A)



(B)

Figure 6.8: OntoFox demonstration of the includeComputedIntermediates setting. Terms that are common ancestors (e.g., Bovidae) to at least 2 external terms are kept in the resulting hierarchy, in addition to the terms (e.g., Primates) explicitly requested.

Each of these six intermediate terms (e.g., Euarchontoglires) is the immediate parent class for at least two child terms (e.g., Primates and Homo sapiens). Primates and mammals are not leaf nodes in the taxonomy hierarchy when the sole parent term is Homo sapiens. Since these terms should be included as well in the final result of the OntoFox output file, they were intentionally included them as low-level source terms. A third choice for including selected terms is inspired by structural modularization techniques. Given a set of signature terms, OntoFox retrieves restrictions that are parent classes of a term. This choice is implemented using OntoFox's SPARQL-based related term retrieval algorithm (Figure 6.9).

Where a restriction mentions another class, restrictions on that class are queried, and so on, until a fixed point is reached. The method gives useful results with the ontologies at the typical level of complexity encountered. It also has the benefit of being straightforwardly implemented in SPARQL and is highly scalable - current modularization algorithms use in-memory representations that require excessive memory for ontologies such as NCBI Taxonomy. Within the OntoFox user interface, users select this choice by choosing "includeAllAxioms". To test OntoFox's SPARQL method to retrieve related terms, three sets of signature terms (individual term, small subset of terms, larger ontology file) were given as input to the OntoFox method and OWLAPI modular-

```
input: source URIs Su, source ontology Oa, SPARQL endpoint se
output: a string saved as an owl (rdf/xml) file
set Sr = empty set, String output=""
while Su is not empty
   pop Ui from Su   #Note: Ui is an item from Su
   put Ui into Sr
 string STRr = get RDF triples of the Concise Bounded Description of Ui
   output += STRr
   set Sp = empty set
   parse RDF triple results and include URIs of referenced entities into Sp
   for each Uj in Sp
      if Uj not in Sr
         put Uj into Su
return output
```

Figure 6.9: OntoFox SPARQL-based algorithm for retrieval of related terms. Its goal is to extract related terms and annotations associated with a set of signature terms (stored in Su) from an external ontology. This method was performed in OntoFox when the setting "includeAllAxioms" is selected.

ization method. In all three cases, both methods generated identical results. One comparative test was performed using the set of terms in the Neurodegenerative Disease Phenotype Ontology (NDPO; `http://ccdb.ucsd.edu/NDPO/1.0/NDPO.owl`) that imports the NIF Lexicon ontology.. The imports closure of this OWL file contains some 50,000 classes in 87 MB of OWL files. Applying the OWLAPI to the classes and object properties in NDPO.owl yielded a module with 1351 classes and 7 object properties - roughly 2.5M OWL file including annotations. The OntoFox generated the same results as measured by the ontology metrics provided by Protégé 4.1. These results support the claim that the OntoFox approach is an effective method for extracting related ontology terms. Finally OntoFox can extract the whole branch ontology terms below a specific ontology term.Choices such as which terms in a parent hierarchy should be included are preliminary to module extraction techniques, which take as input a set of terms (signature) that the ontology developer has identified as being of interest. OntoFox supports experimentation by offering more than one choice for making such a term selection.

OntoFox data input and result output

Besides the web form-based data input, data can be uploaded as a text file to the OntoFox web server. This input file contains the same information as the web form input method, but makes it easier to submit batch jobs. The file upload method also makes it possible to keep track of submissions and easily update the input. An OntoFox sample input file (available at `http://ontofox.hegroup.org/format.txt`) has been developed for users to quickly understand and use the required format. Also, the OntoFox input file can also be automatically generated using the button "Generate OntoFox Input File" from data in the web forms (Figure 6.5). Finally, jobs can be programmatically submitted to the OntoFox server via a script at `http://ontofox.hegroup.org/service.php`. As an example, the following command line can be used to provide an input file (input.txt) and retrieve the corresponding output file (output.owl):

```
curl -s -F file=@/tmp/input.txt -o /tmp/output.owl http://ontofox.hegroup.org/service.php
```

An OntoFox query can result in either a processing error, in which case an explicit message is provided to the user, or in the production of an OWL file serialized in the RDF/XML format. This OWL file constitutes an ontology on its own and can be visualized using the Protégé ontology editor [126] and directly imported into the target ontology using the OWL import directive. The OntoFox process can be executed at different times to import updated information of external ontology terms. By keeping and updating the original OntoFox input text file, users can subsequently query the OntoFox server on a regular basis and get up to date information with little effort.

OntoFox application in Vaccine Ontology (VO) development

Using OntoFox, VO currently imports approximately 1000 terms from 12 external ontologies such as GO [53], NCBI Taxonomy, OBI, PATO [128], and Mammalian Phenotype Ontology (MP) [63](Table 6.2). When using OntoFox to develop VO, it was desirable to apply different settings depending on the source ontology considered, and therefore generated one OWL file to be imported per external resource. Once imported into VO, external terms can be used exactly in the same way as other vaccine-specific VO terms. Different OntoFox settings have been applied for generating these 12 ontology subsets for VO imports (Table 6.2). In terms of superclass extraction, six were generated with the OntoFox setting "includeNoIntermediates", which is particularly useful when the intermediate superclasses do not generate much more information needed for the target ontology. The setting "includeComputedIntermediates" was used for extracting ontology terms

from three external resources, including NCBITaxon, PATO, and the PRO [31]. In the case of the NCBI taxonomy it reduces the number of imported classes without losing the information of the most recent ancestor superclasses (Figure 6.8). Finally, the setting "includeAllIntermediates" has been used for extraction from OBI, ro_proposed (`http://purl.org/obo/owl/ro_proposed`), and the Sequence Ontology (SO) - `http://www.sequenceontology.org/`) (Table 6.2). These three external ontologies are closely related to VO, and their original hierarchies should be maintained for those terms imported to VO. Similarly, different annotation property settings have been applied (Table 6.2). Typically, VO follows the IAO's ontology metadata scheme and uses the properties "rdfs:label" or "iao:definition". To make the annotation styles consistent among all ontology terms in VO, the OntoFox directives "copyTo" and "mapTo" were used (Table 6.2).

6.4.4 Discussion

While an implementation of the MIREOT strategy has been performed in the context of OBI, it relies on command line scripts, making its use impractical for the average ontology curator and limiting its adoption by interested users. Comparatively, OntoFox provides a convenient web-based approach to use MIREOT that does not require programmatic skills and allows users to specify their requirements via simple text formats. In addition, the OntoFox server provides additional options for users to add and rewrite annotations, to include superclasses or subclasses, or select terms via related restrictions (transitively). This last option performs comparable to existing structural modularization methods. OntoFox uses a RDF triple store and SPARQL for information storage and retrieval, resulting in a system that scales better than in-memory modularization techniques. For the Neurodegenerative Disease Phenotype Ontology, OntoFox extracted the same module that a more sophisticated modularization technique did. While these more sophisticated techniques may be desirable, there are issues with their use. While OntoFox uses simpler methods to retrieve terms and axioms related to MIREOT specified terms, it provides a simpler and more understandable approach to reuse. This is particularly useful in conjunction with the fact that OntoFox provides an easy approach to incorporate frequent updates from source ontologies that are under active development. The provision of a simple mechanism for importing selected terms from external ontologies does not shield the user from general issues associated with using external terms. When using terms from other ontologies, care must be taken to avoid a situation in which the meaning of an ontology term in the source ontology is different from the meaning of the term used in the target ontology. To avoid this problem, users are advised to exercise due diligence when selecting

terms to import. OntoFox helps prevent this confusion, by first offering a limited set of 15 selected ontologies with good documentation and second by importing annotation properties, providing immediate access to the textual definitions. Where an ontology developer has questions as to the meaning of a term it is recommended that they contact the developers of the source ontology and ask for clarification and enhanced documentation. The 15 initially selected ontologies generally have trackers and mailing lists where questions can be posted. Another issue is the evolution problem associated with using ontologies that are under active development, as is the case with most current biomedical ontologies. Although at a certain time point a certain term is used in the source and target ontologies equivalently, over time the usage of the term (and the associated classes) in both ontologies may change. It is considered good practice to not use terms from external ontologies in ways not consistent with their definition, and for ontologies to deprecate old and define new terms rather than changing the meaning of terms. While OntoFox provides a way for users to automatically update the annotations of imported terms, it cannot monitor changes in meaning. Therefore it is up to developers to choose ontologies that have practice that will let them monitor for such changes and make adjustments as appropriate. OntoFox's 15 initial ontologies were chosen because they tend to have predictable practices related to ontology evolution.

6.5 Conclusion

The common ID policy has been adopted has a normative principle for Foundry resources within the OBO Consortium, and it is expected that OBO library resources will abide by it. One strong incentive for developers to do so was coupling the Ontobee dereferencing service with the obtention of the common prefix and using OBO types of PURLs. The implementation of the Ontobee is described in the following chapter, Chapter 7. The IAO common metadata set is being used by multiple ontologies: it provides the common annotations supported by OntoFox and Ontobee. Work is in progress to augment it with the annotation properties required to enable automated OBO to OWL conversion[21].

While the current implementation of the MIREOT mechanism is tailored towards OWL ontologies, a similar mechanism could be applied to OBO format resources. It is also expected that an option in the released version of ontologies, such as OBI, will in the future enable the replacement of *external.owl* with *imports.owl*, a file of imports statements generated by extracting the ontology

[21] `http://oboformat.googlecode.com/svn/trunk/doc/obo-syntax.html`

URIs mentioned in *external.owl*. Users would then be able to import all of the external resources, therefore replacing the MIREOT selected terms.

In the case of OntoFox, more ontologies will be included in the list of source ontology repositories. These ontologies may come from the OBO foundry or other reliable sources. Developing an OntoFox plugin for ontology editors (e.g., Protégé) is also under consideration. Editors of OBO format ontologies desire a similar facility, and while OntoFox currently supports resources in OWL, integrating an automatic conversion for OBO files could directly support the OBO format. As usability testing of the web interface has thus far been informal, more careful usability studies will be designed, such as a survey to solicit feedback from the community. A drawback to the current OntoFox approach is that it requires maintaining independent text files with the import directives (compared to the original MIREOT mechanism which reads in an existing OWL file). However, given that there are no editing tools available in either case, the human readable format seems adapted.

Finally, as module extraction technology matures, the ability to use such mechanisms for doing targeted imports, on a source-by-source basis, will be included.

Table 6.1: The 15 source ontologies currently available in OntoFox

Ontology	Source ontology URI	Example of source Ontology Term URI
CARO	http://purl.org/obo/owl/CARO	http://purl.org/obo/owl/CARO#CARO_0000040
CHEBI	http://purl.org/obo/owl/CHEBI	http://purl.org/obo/owl/CHEBI#CHEBI_48999
CL	http://purl.org/obo/owl/CL	http://purl.org/obo/owl/CL#CL_0000799
DOID	http://purl.org/obo/owl/DOID	http://purl.org/obo/owl/DOID#DOID_12685
ENVO	http://purl.org/obo/owl/ENVO	http://purl.org/obo/owl/ENVO#ENVO_00000483
FMA	http://purl.org/obo/owl/FMA	http://purl.org/obo/owl/FMA#FMA_9712
GO	http://purl.org/obo/owl/GO	http://purl.org/obo/owl/GO#GO_0043152
IDO	http://purl.obolibrary.org/obo/ido.owl	http://purl.obolibrary.org/obo/IDO_0000064
MP	http://purl.org/obo/owl/MP	http://purl.org/obo/owl/MP#MP_0000026
NCBITaxon	http://purl.org/obo/owl/NCBITaxon	http://purl.org/obo/owl/NCBITaxon#NCBITaxon_263
OBI	http://purl.obolibrary.org/obo/obi.owl	http://purl.obolibrary.org/obo/OBI_0100026
PATO	http://purl.org/obo/owl/PATO	http://purl.org/obo/owl/PATO#PATO_0001793
PRO	http://purl.org/obo/owl/PRO	http://purl.org/obo/owl/PRO#PRO_000001795
SO	http://purl.org/obo/owl/SO	http://purl.org/obo/owl/SO#SO_0001288
VO	http://purl.obolibrary.org/obo/vo.owl	http://purl.obolibrary.org/obo/VO_0000001

Table 6.2: OntoFoxed ontologies in VO

	Ontology Name	# of signature terms	# of imported terms	Intermediates	Annotations
1	CARO	2	2	No	rdfs:label *copyTo* iao:preferredTerm
2	CHEBI	13	13	No	oboInOwl:hasDefinition *copyTo* iao:definition
3	DOID	10	57	All	oboInOwl:hasSynonym *mapTo* iao:alternativeTerm
4	FMA	2	2	No	
5	GO	2	2	No	
6	IDO	1	2	No	
7	NCBITaxon	143	198	Computed	
8	OBI	41	48	All	rdfs:label, iao:definition
9	PATO	15	17	Computed	rdfs:label *copyTo* iao:preferredTerm
10	PRO	2	2	Computed	oboInOwl:hasDefinition *copyTo* iao:definition
11	ro_proposed	7	9	All	
12	SO	1	1	No	oboInOwl:hasSynonym *mapTo* iao:alternativeTerm

Chapter 7

Publishing biomedical resources on the Semantic Web

7.1 Introduction

The goals of the OBO Foundry, and of biomedical ontology in general, are very much in line with those of the semantic web, and using semantic web technologies for using and sharing biomedical data annotated with ontology terms is desired. One such technology, now existing in a number of implementations due to the popularity of LOD, is the practice around serving linked data so that it can be browsed and accessed at the granularity of instances [185]. While such services also somewhat address serving ontology terms and relations as well, existing implementations did not satisfy current needs. For example, I am not aware of a linked data service that accurately renders logical axioms, such as property restrictions, expressed in OWL. To make ontology terms accessible there is a need to, as with linked data, (1) provide human users with adequate information to understand what the term means, while also (2) make available the documentation of these terms in a form that automated tools can use. Each of these goals is here modified from the case where typically instances are browsed. The human browsable presentation of ontology terms is intended for a few key audiences. The biomedical community is a diverse community with different practices and perspectives, and frequently uses different words to describe the same types of entity. Ontology developers are responsible for building ontologies. They need to be able to easily navigate their own work as well as the content of other ontologies in order to be able to find existing terms (and confirm they are indeed relevant) that should be reused rather than created de novo. Curators and annotators work with existing datasets and must be able to discover and understand the terms applicable to their data.

7.1.1 Requirements

Experience and iterations of discussions have led me (in collaboration with Alan Ruttenberg) to the summarization of two sets of requirements, one aimed at providing a useful user experience (U1-U9) and a second set of goals related to engineering (E1-E6).

Based on my experience within the community of practice, and after discussion with prospective users, the requirements, aimed at providing a useful user experience, are to:

U1) Provide a service with predictable behaviour across the whole body of OBO ontologies.

U2) Ensure that useful information is displayed. This includes at a minimum documentation, attribution, and provenance. Terms should be displayed with labels rather than identifiers, but identifiers should be accessible.

U3) Be clear as to what IRIs identify.

U4) Term IRIs should be used in scholarly citations. Common ways of bookmarking should yield the term IRI.

U5) Deliver RDF that is accurate when compared to the source ontology. Ontology writers use specific relations and axioms to communicate and set expectations that users of their work will be able to retrieve them as they were written.

U6) Present both ontology-centered and term-centered views. Access will be via two common routes: either start at a given ontology and explore or search within it, or directly request the page for a term via its IRI or a search result.

U7) Display OWL expressions in a readable syntax. The RDF-centric rendering of OWL is difficult to understand beyond simple statements.

U8) Be able to customize views as desired by the ontology developers. Often ontologies, such as UniProt [97], have web sites presenting their work, and some have term browsers. Their developers don't want to lose their "branding" by having a different site be the destination for viewing terms but at the same time wish to take advantage of the services Ontobee provides.

U9) Provide tools to aid navigation to ontology terms of interest. For example, any ontology term appearing anywhere on the page should be clickable to view information about it, and the ability to search for terms should always be easily available. While ontology navigation

and visualization is an ongoing area of research, efforts have been made to incorporate what is known, as well as users' feedback, into this work.

On the engineering side, there is a different set of requirements, motivated by the desire to take advantage of already widely available semantic web technologies as well as promote their uptake in the community.

E1) Adherence to documented web and semantic web practices. The relevant specifications are for RDF [91], RDF/XML [181], OWL 2 [186], SPARQL [187], and XSLT [188]. Using these standards means there are a number of implementations to chose from, providing advantage of advances in performance and functionality without changing the underlying code base.

E2) Predictable access to RDF/XML assertions relevant to computing with the term. In order to ensure predictability advertised policies that let developers rely on what they will be able to get to are needed.

E3) Ability to have generated HTML reused by other applications.

E4) Visibility in search results of popular search engines

E5) Scalability as the number of terms served increases, and as the number of clients increase.

E6) Transparency. For example, an interested user should be able to see the queries that are used to collect the information that is assembled into the web presentation.

Finally, a general requirement of this work is that it be built on an open source platform. By doing so collaborative development, extensions, and experimental forks by others can happen.

7.1.2 Previous work

There are a variety of LOD and ontology browsers that have been developed. The NCBO Bioportal [189], Ontology Lookup Service (OLS) [49], Manchester Ontology Browser [190], and AmiGO [135] primarily offer views and navigation of biomedical ontologies without particular attention to operating in the framework of LOD. Ontotext's Linked Life Data [191] and the Bio2RDF effort [66] are projects that are closest to Ontobee. DBpedia [192] is an exemplar of LOD most associated with the movement, and are both primarily instance oriented. DBpedia serves linked data derived from Wikipedia and Virtuoso's data spaces aggregate many different sources of data and present them as linked data.

The Ontology Lookup Service (OLS) provides both a web-based interface and a programmer's API based on SOAP. The web-based search interface provides autocompletion for terms in OBO format ontologies and presents three different views, one of the term alone, one of the term in hierarchical context, and one graphical visualization of the term view with either ancestors or descendants [193]. OLS differs in that it doesn't support parsing of resources in the Web Ontology language (OWL) and does not provide RDF/XML format data (requirement E1). Instead there is the SOAP based API that provides access to search facilities, terms, and relations between terms and other terms. Still, the OLS is appreciated within a large community of biomedical annotators. Of note is the inclusion of clickable graphical display of either the path from the term to the ontology root, or of children of the term.

The BioPortal provides a variety of services including textual and graphical term browsing and search [189], as well as REST-based APIs for accessing and using ontology terms, including several that deliver RDF. Textual presentation of ontology terms is only partially covered - classes are browsable, but not relations or instances. Of note, textual views do not satisfy U7 in that property names and axioms that are displayed are not hyperlinked to pages that describe the terms. It only trivially satisfies U5, in that it omits logical axioms. This can be demonstrated by comparing the display of the OBI term OBI_0001705 in BioPortal and Ontobee [194]. Although there is an RDF service it is not coupled with the user-oriented interface as it would for a typical LOD browser, and it fails U3. RDF retrieved by the API call does not contain all the axioms from the original ontology, and rewriting the OWL changes some assertions on some properties to assertions on different ones, typically from the SKOS vocabulary [195]. In addition the BioPortal has recently provided a RDF triple store and SPARQL query browser [196]. However the RDF generated, similar to the service, does not satisfy U3 in that is also rewrites certain properties and does not provide a way to get all the assertions for a term. It also skolemizes blank nodes, yielding, in the case of SPARQL CONSTRUCT queries, RDF/XML that will be rejected as invalid by tools that parse OWL [197].

The Manchester Ontology-browser [190] is intended as an on-line ontology browser but not specifically as linked data browser. It focuses on the accurate and accessible display of fully reasoned-over OWL 2 ontology content including logical axioms, but does not serve RDF for individual terms. It is notable for clear display of owl axioms, consistent use of labels, and for presenting the reasoned over ontology so that consequences beyond the asserted axioms can be understood by the user. The Manchester Ontology Browser is an ontology-centric interface. Access is typically via ontology IRI followed by search or navigation to a term and IRIs displayed in the address bar are

not the term's IRI nor persistence, though some element of persistence is offered via permalinks.

AmiGO [135], another example of a web-based ontology browser, in this case developed by the Gene Ontology Consortium [198] in order to browse the Gene Ontology. AmiGO has been developed over time to meet their community (typically biologists) needs. It provides search by label, various hierarchical displays, including some generated by simple inference, and display of gene products that are annotated by the class in focus. However with respect to existing requirements AmiGO has several shortages. It does not serve RDF, and so does not operate as a linked data server. It is oriented towards the display of OBO format assertions, which do not always easily map to the OWL equivalents, which are not displayed. It is notable, as it is the best example known of an ontology browser that serves the needs of a specific community.

Ontotext's linked life browser [199] provides a web page interface as well as the ability to retrieve RDF for a term as RDF/XML and several other RDF syntaxes. While formats are offered as link with distinct IRIs, content negotiation is also active. However HttpRange-14 is not followed - leading to a failure of U3. As with the Bioportal and Bio2RDF, the standard IRIs for OBO terms, such as for terms from the Gene Ontology [191] are not used, and the assertions are adaptations of the original assertions expressed using the SKOS vocabulary. Somewhat confusing is the handling of subclass assertions. In the default web view, and in the RDF, they are simply omitted. However an option on the web page under the title "inference", when chosen, shows the existence of skos:broaderTransitive relations in place of subClassOf. Even with this setting, RDF retrieved using the links or with content negotiation does not contain these relations, which, even in translated form, are an essential feature of the GO. Thus U2 and U5 suffer. Bare URIs are displayed when objects of predicates. Search is via keyword, and is uncomfortable in that a limited number of results are shown and there appears to not be any sorting for relevance. This is exemplified by a search for "biological process", an upper level term from the GO that nonetheless does not appear on the first few pages of search results.

Bio2RDF [66], is another effort to create a linked data view of both biomedical ontologies and databases. In what seems to be the common pattern, it issues new IRIs for existing resources and rewrites those resources according to another schema. Display of terms varies. Version 1 resources use the Pubby [200] software for term display, generally favoring IRIs over rdfs:labels for display, and with no provision for displaying OWL axioms other than as raw triples. For example, `http://bio2rdf.org/page/go:0032283` shows the target of some of the subClassOf relations as "(Unnamed RDF node)" rather than a readable OWL restriction [201]. Provenance for terms is

often absent. Whereas for a GO term such as go:0032283 there is a dc:license link that can be interpreted as provenance, the the term 'scalar measurement datum' [202] from the Information Artifact Ontology, authored natively in OWL 2, is rendered as `http://bio2rdf.org/page/iao:0000032` [203]. In that rendering none of the axioms are displayed, labels are not used, nor even visible as a property value and there is no indication from where the single triple displayed originates. RDF/XML is available via hyperlink, and by content negotiation, and matches the html display. The BIO2RDF Release 2 resources are presented as web pages using Openlink's Virtuoso Faceted Browser. An example of term display for a term in NCBI taxonomy, 'bos taurus' [204] is [205]. Here the view is notable for the combination of bare IRIs used as property labels with a fixed width display. This yields IRIs in which the middle part has been replaced with an ellipsis, for example `http://bio2rdf.org...bulary:name_class`. The link from the title of the page actually resolves to the Pubby display, which is not completely concordant with the original page, contributing to further confusion.

DBpedia extracts structured information from Wikipedia and to make this information available on the Web [192]. DBpedia uses content negotiation to return RDF descriptions when accessed by Semantic Web agents and a HTML view of the same information to Web browsers. The HTML web browser does not provide hierarchical tree structure. DBpedia is focused on linking instance data (mostly outside life science domain) instead of ontology terms. The DBpedia Ontology has been developed to support data linkage and mapping within the DBpedia datasets [192].

In this work those issues are addressed, implementing a service that provides a balance between following Web and Semantic Web specifications while being careful about ontological issues as detailed below.

7.2 Implementation

7.2.1 Overview

The Ontobee server is currently a single HP server running Red Hat Linux operating system (Red Hat Enterprise Linux 5 server). The open source Apache HTTP Server is used. Programming is done with PHP, Java, SPARQL 1.0, and JavaScript. OpenLink software's Virtuoso Open-Source Edition is used as a RDF triple store. The same machine both runs the triple store and generates the documents needed to implement the web interface.

As shown in Figure 7.1, Ontobee provides access to RDF and HTML documents with informa-

tion for ontology terms. RDF documents can be accessed at the published identifier for the term implementing the "httpRange-14" recipe of issuing an HTTP redirect to a document after first responding with a 303 status code [206]. To provide a user-friendly web interface for users to identify related links and detailed information, the RDF document includes a stylesheet directive [188], which the browser uses to generate HTML. In addition direct access to HTML is provided at a different IRI as shown below.

It is not uncommon for there to be different versions of an ontology available. This leads to a design choice regarding what happens when different ontologies loaded into Ontobee import different versions of the same ontology - should the most recent version be loaded and all imports modified to use that version? Or should each ontology and its imports be kept segregated so that different versions for different imports are supported? For now, Ontobee chooses the latter strategy, using a separate named graph for each top-level ontology it loads. While it prevents issues that could be created by "forcing" a resource to use a newer version of an imported ontology, it means that on the same server multiple versions of the same resource are available, which may be confusing for users.

7.2.2 Access to descriptions of entities referred to by term IRIs

A common method by which RDF and web pages are associated with a term is to have the term IRI accessed via a server configured to do content negotiation. In that scenario, the user agent issues a GET with the term IRI, and sets the HTTP Accept header with the mime-type desired, typically application/rdf+xml or, for the web browser, text/html. Content negotiation was forgone in favor of using the mechanism described in the resolution of httpRange-14, under the rationale that

> a) Documents are in the domain of discourse of our field, and returning different documents in response to access requests for a single IRI is confusing because there is confusion about what the IRI denotes. For example, in working with a commonly available database such as one at NCBI, one might want to identify the class of protein properties of its instances, or the evolution of the web page that is displayed when the IRI is dereferenced in the browser, or run a processing pipeline on the information about the protein class formatted in XML. If all three are given the same IRI it is difficult to record assertions specifically about one or the other. However if each has a distinct IRI one or the other can be referred to by using the

appropriate IRI.

b) Doing so promotes predictability. Extant servers vary on their implementation of Accept header processing, commonly returning content types other than what is requested.

c) It is not always possible to easily set Accept headers in requests. For example few web browsers offer the ability to change the Accept header. Programming APIs may or may not expose this functionality. With APIs that do allow this, documentation can be buried in details that are easy to miss.

d) httpRange-14 offers a solution that is simple and uniform

The W3C Technical Architecture Group (TAG) resolved issue httpRange-14 by saying the following:

> If an "http" resource responds to a GET request with a 2xx response, then the resource identified by that URI is an information resource;
>
> If an "http" resource responds to a GET request with a 303 (See Other) response, then the resource identified by that URI could be any resource;
>
> If an "http" resource responds to a GET request with a 4xx (error) response, then the nature of the resource is unknown.

The server architecture implements this resolution by having all HTTP access requests for entities named in an ontology to result in a response status code of 303. By doing so, any potentially confusing implications that the entity is an "information resource" are avoided. The 303 response also includes a redirect to Ontobee, which provides an RDF/XML document describing the entity. This is the middle case above in the httpRange-14 resolution.

Following the redirect, the client requests the RDF/XML document, which has its own IRI, and the server responds with a 200 status, indicating an information resource as per the httpRange-14 resolution.

7.2.3 Use of PURLs

Within the OBO community, the common (and encouraged) practice is to create PURLs for ontology ids using the purl.obolibrary.org domain. This facilitates changing the servers used to respond

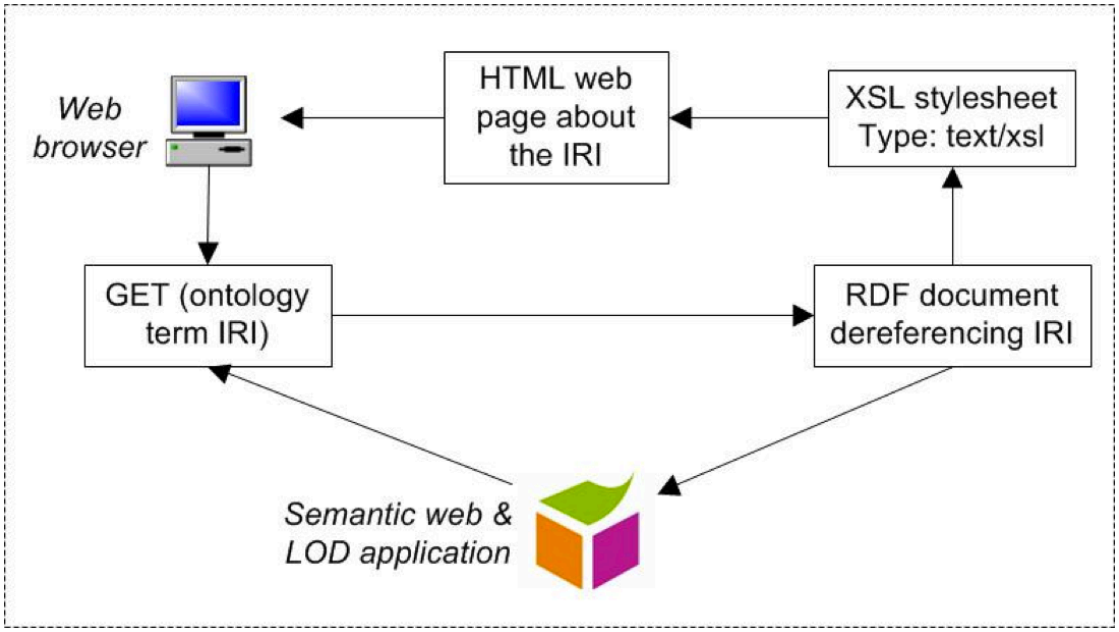


Figure 7.1: Ontobee system architecture design. Queries for an ontology IRI term from a web browser or a http user for a Semantic Web and LOD application will send a GET (ontology term IRI) request to the Ontobee server. Once a request is received, the Ontobee server will issue a SPARQL query against a RDF triple store and return a RDF document that dereferences the IRI. The RDF document will be returned to the Semantic Web and LOD application (no web browser involvement). For a web browser, the browser notices the XSL stylesheet and asks for that. The XSL stylesheet returns the HTML with an XSLT wrapper. The browser applies that to the RDF, gets the HTML and renders it.

to access requests, without having to change the IRI of the term. Where to redirect PURLs is a choice of ontology developers. While the use of Ontobee is recommended, there is also the option for custom redirection. In order to reduce the cost of administering a PURL server, the Online Computer Library Center (OCLC) granted permission to set up a Canonical Name Record (CNAME) for purl.obolibrary.org, making it an alias for purl.oclc.org. This allows us to leverage existing infrastructure - OCLC's PURL resolver - while providing a backup mechanism should OCLC's server stop being available. In that case the DNS entry for purl.obolibrary.org could still be redirected to a different PURL server or implementation.

7.2.4 Ontology retrieval and preprocessing

Ontobee retrieves, pre-processes, and loads ontologies into its triple store. The server currently loads OBO Foundry ontologies as well as a selection from the OBO Library [207]. These ontologies are distributed in either OBO format or in OWL. The obo2owl pipeline [208] is used to generate an OWL version of OBO ontologies according to the translation for OBO Format 1.4 [209].

A set of PHP/Java scripts retrieves the OBO library ontologies on a regular basis (currently daily). Ontobee does no OWL reasoning on the source ontologies, though some developers release a pre-reasoned version including inferred axioms.

7.2.5 Retrieval of information about a term

SPARQL was used to retrieve information from the triple store. Many of the queries are SPARQL 'describe' and 'select' functions against the RDF store. The information retrieved includes:

- Annotations on the term
- Restriction superclasses of the term, when a class is specified equivalentClass assertions where the term is one of the equivalents
- The ancestors of the term
- The direct instances of the term
- Other terms in the ontology that reference the term in their axioms
- Other ontologies, of the ones in Ontobee, that also reference the term
- Ontology header and ontology annotations
- The SPARQL queries used for retrieval

The motivation behind using the above information should for the most part be obvious. However some items deserve mentioning. In particular, the ontology annotations are queried so that attribution information, such as Dublin Core (dc) metadata values of dc:contributor and dc:creator, can be known to a casual user who accesses only one or a handful of terms. The SPARQL queries are collected so that they can be shown to users of the web interface who wish to learn more about how to use semantic web technologies.

A variety of choices have been made to decrease the number of queries executed, reduce the size of and increase the readability of the resulting RDF. An example is the use of 'transitive' and the 'CBD' (i.e., Concise Bounded Description [210, 211] options in the Virtuoso SPARQL for some queries. For example, in order to retrieve the ancestors of a term, one might have to issue a number of separate queries as the class tree is traversed. Use of the 'transitive' option in Virtuoso's implementation of SPARQL obviates that need. Retrieving the axioms for a term can also take a number of SPARQL select queries.

The current approach to retrieval effectively offers a compromise between expressivity and query execution time. Therefore the set of assertions about a term is not necessarily complete, but generally shows enough to be useful. For example if the loaded ontology does not have inferred superclasses, such as those in an equivalentClass axiom, these would not be retrieved. It is expected that SPARQL implementations over OWL will improve in time and the technology used will be reexamined periodically.

To minimize the size and increase legibility of the RDF/XML document, the raw output from SPARQL queries is reformatted. For example, automatically generated namespace prefixes (e.g., xmlns:n0pred) are rewritten using more widely used ones (e.g., xmlns:obo), and the RDF is re-rendered using the OWLAPI [150].

7.2.6 Generation of RDF and HTML outputs

Upon request for an ontology term IRI, Ontobee generates a representation of that term in RDF. The RDF is returned, in RDF/XML format, including a stylesheet directive. When the request is from a web browser agent, the stylesheet is retrieved. While this stylesheet could programmatically transform the generated RDF, it is easier to generate the HTML in a more fully featured programming language. Therefore the HTML is generated using PHP and then encapsulated within a trivial XSLT transformation that translates the root RDF node into the resultant HTML. This HTML is subsequently used by the browser to render the page. In addition to the term IRI, these generated documents are given their own IRIs and are accessible from the server independently. Although they may not be of primary interest to a user browsing the site, this access was provided should it be useful to retrieve them directly in developing another application or in order to make assertions about them, for example in a study of how the term's logical definition has changed over time. As a final touch, the RDF is made valid OWL by adding an ontology header and imports statements from the original ontology. Lightweight linked data clients will ignore these statements,

but doing this make it possible to open the RDF document in an OWL-aware tool and give the correct inferences should the reasoner be employed.

7.2.7 Search

A keyword-based search facility with autocompletion is available. On the client side, autocompletion is implemented using the jQuery JavaScript Library [212] and AJAX [213]. Users can make calls to the server which implements the search. Search over the entire set of ontologies is available at `http://ontobee.org`. Each ontology term page also has a search box that is scoped over terms from the same ontology.

7.3 Results

Ontobee was initially deployed in late 2009. Near the beginning for 2012 it became the default location for the bulk of OBO ontologies listed at the OBO Foundry web site. Ontobee undergoes continuing development as suggestions are received and usage evolve. Source code for the server can be found in the subversion repository hosted by Sourceforge at `http://sourceforge.net/p/ontobee/code/`, and is distributed under the Apache License, Version 2.0. Below the current web interface, scale, and adoption are discussed.

7.3.1 RDF

Figure 7.2 shows a portion of the RDF/XML generated for the term 'vaccine' from the Vaccine Ontology, `http://purl.obolibrary.org/obo/VO_0000001`, as seen when one chooses view source in a web browser. This file is a valid OWL-DL document - a small ontology centered around a single term - made so by the addition of the ontology header and import statement. As a result users can open an individual term IRI directly in an OWL editor such as Protégé [126], and run a reasoner. Reformatting with OWL API provides indentation to make logical definitions easier to read. Not shown, the RDF includes further information that was judged to be of utility in a linked-data context, such as attributions information about the ontology and labels for any term used in the RDF. The precise contents of the RDF are evolving as experience grows. For example, in the future, a modularization algorithm may be used to select relevant assertions. The RDF and HTML documents are accessible separately. For the class 'vaccine', the IRIs would be:
`http://purl.obolibrary.org/obo/VO_0000001` denoting the class,

`http://purl.obolibrary.org/obo/vo/about/VO_0000001` denoting the RDF/XML document, `http://purl.obolibrary.org/obo/vo/page/VO_0000001` denoting the html document.

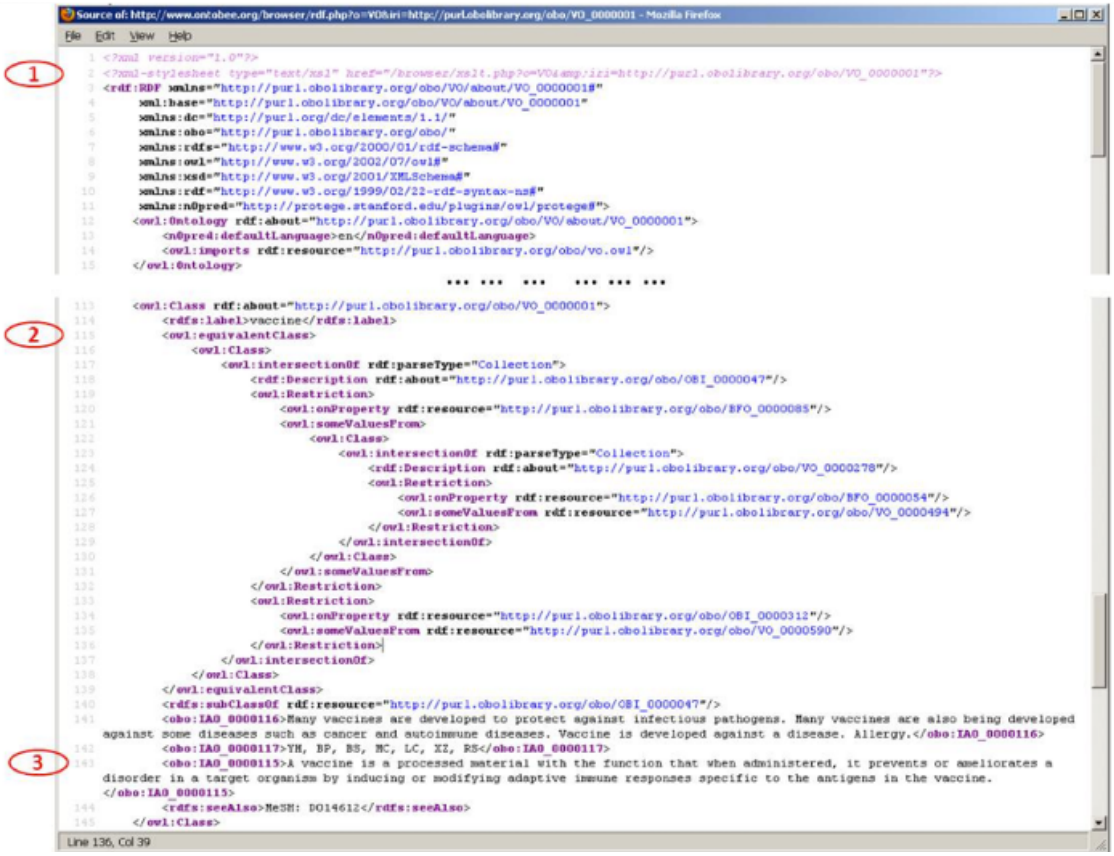


Figure 7.2: An Ontobee RDF output file, which is the source page of the ontology term `http://purl.obolibrary.org/obo/VO_0000001` (label: 'vaccine' from VO). The RDF document includes an XSL stylesheet directive (1). (2) Shows part of the logical definition, in this case, an equivalent class axiom. (3) text definition.

7.3.2 Web interface

Figure 7.3 shows the same term, 'vaccine', in the web interface. The items and order of items have been chosen to give ample information to help the user understand the term, and to emphasize elements that contribute to reuse. rdfs:labels, when present, are used for display. To encourage discovery, any visual element that is an ontology term is a link to the page for that term. Below is a description of the elements of the page and motivation for their placement and inclusion.

At the top there are the type of term, its definition, and the term IRI, bolded to emphasize that it is what the user should copy should they want to cite the term. All annotations on the term,

such as editor notes and synonyms, and term editor are just underneath. After the definition, these are the most understandable documentation. Next are equivalents, the strongest form of logical definition as they are both necessary and sufficient conditions. Here and elsewhere the display is formatted using a variant of Manchester Syntax in which term labels are displayed instead of IRIs. The hierarchical context gives the user more information about the term and draws their attention to more general and more specific terms that may be appropriate for their work. Direct superclasses and class axioms follow. These refine the logical definition, provide links to related terms, and serve a pedagogical role by presenting patterns that might be used for other definitions. Other terms in the ontology whose axioms make some reference to this page's term are next. These axioms give more information about how the term can be used in practice as well as help understand the term by seeing it in use. Other ontologies within Ontobee that use the ontology term are then displayed. Reuse is encouraged, and showing that there is already use in other ontologies lets the user know where this is happening as well as navigate to the other ontology for further examples of using the term. Finally, there is an offer to show the SPARQL queries used to generate the page. Users with technical experience are encouraged to adopt SPARQL and other semantic web ontologies. By exposing the queries they can learn more about the technology and try it out by themselves.

7.3.3 Search

Ontobee provides a simple textual search (Figure 7.4). On the Ontobee front page, one can query ontology terms across all ontologies. When viewing an ontology page, or term from an ontology, the search is across terms in that ontology and its imports. As the user types, commencing on the third character, a drop-down menu with terms whose label contains the string typed so far (Fig. 7.4A). A selection of a specific term in the drop-down menu will lead to the web page dereferencing the ontology term (Fig. 7.4B). Selecting one of the menu items navigates to the page for that term. Alternatively one can choose "Search terms" to get a page that lists all matches, sorted in order to first show terms that start with the search string, shortest to longest, then terms that include the string, shortest to longest (Fig. 7.4C).

7.3.4 Scalability

Ontobee scales well in practice. Currently it provides access to over 1,300,000 ontology terms from more than 100 ontologies without appreciable delay, including very large resources such as the NCBI Taxonomy. Table 7.1 shows a selection of these ontologies. Scalability is achieved in

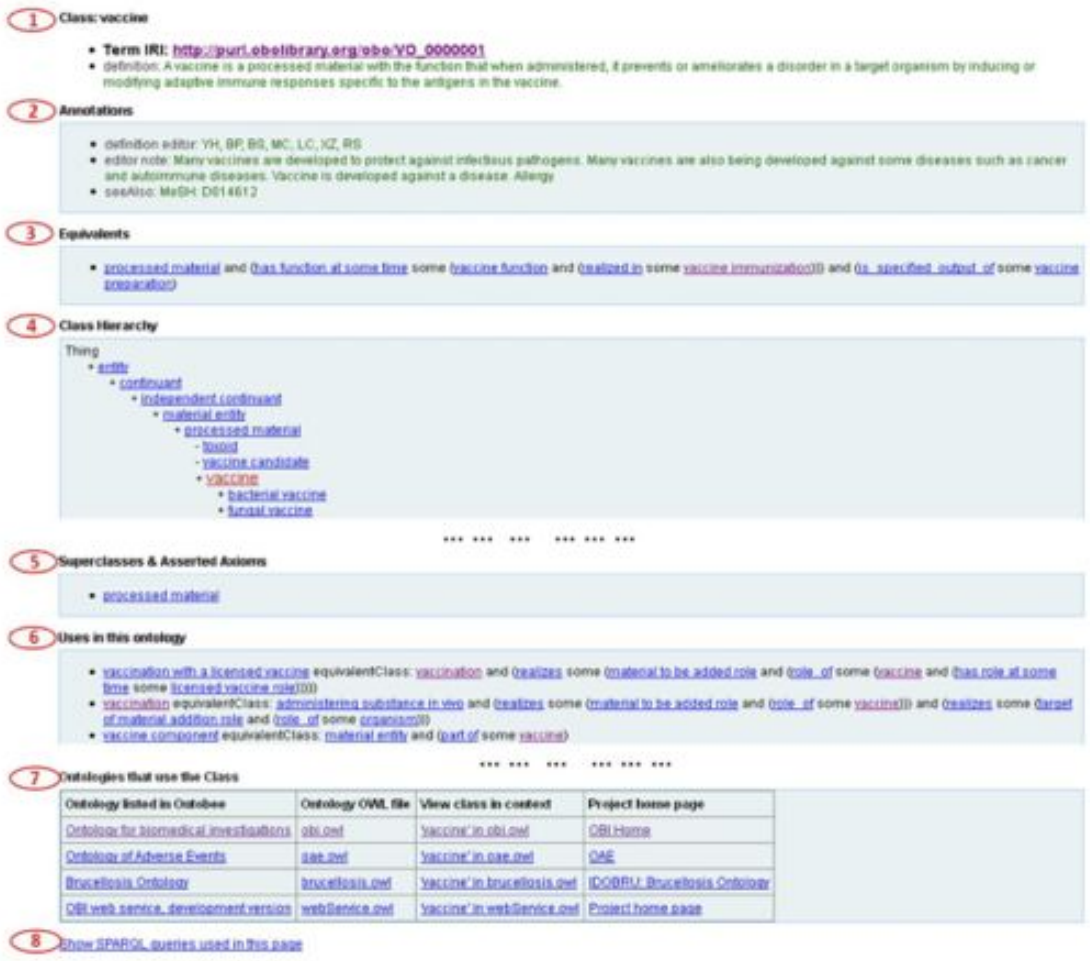


Figure 7.3: Ontobee HTML rendering of the VO term 'vaccine' as seen in Firefox.

large part by using a triple store and SPARQL. Performance of triple stores and SPARQL is an active area of research, has been improving over the last few years and it is expected to continue to improve. Since the common SPARQL technology is used for querying and web page displaying, different triple store implementations can be tested as the technology or needs evolve.

7.3.5 Community adoption

Ontobee was initially prototyped with the VO and OBI ontology groups. Over the last year it has begun to serve as the default destination for IRIs of terms in ontologies listed on the OBO Foundry site.

Google Analytics data shows that the number of Ontobee daily users has steadily increased since early 2012 (Figure 7.5). During the year of 2012, there were over 10,000 unique visitors. On

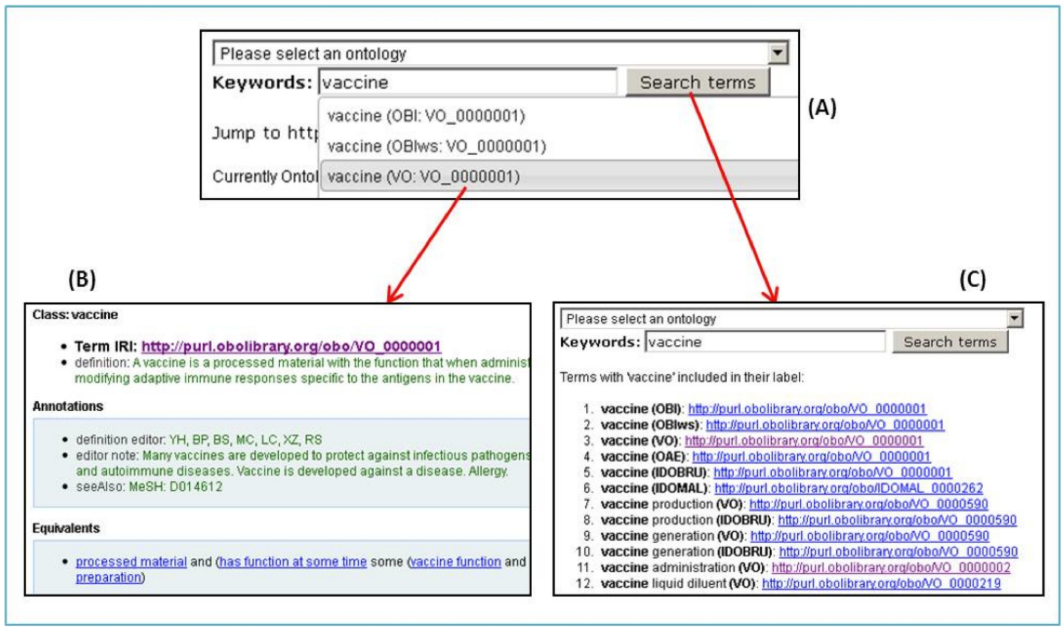


Figure 7.4: Demonstration of the search capability and the HTML rendering of the VO term 'vaccine' in the Firefox Ontobee web browser. (A) On the Ontobee home page, a search of 'vaccine' returns a number of results from different ontologies. (B) Selecting the VO term 'vaccine' from the search result list navigates to the page for 'vaccine' in VO term. (C) A click on the "Search terms" button results in the display of all links for all terms with labels that match 'vaccine' in Ontobee.

average, each visitor spent nearly 4 minutes on the site and browsed about 4 pages.

A variety of projects use Ontobee as their preferred term visualization tool. Eagle-i [153] is one such project, the result of an NIH-funded effort to help scientists discover research resources. Eagle-i provides access to information about more than 50,000 resources and is deployed at a growing network of universities.

Table 7.1: Summary of selected ontologies available in Ontobee. The number of terms includes terms defined inside the ontology as well as terms imported from other ontologies.

Ontology Name	Number of terms
GO (Gene Ontology)	38,562
PR (Protein Ontology)	35,342
PATO (Phenotype Ontology)	2,331
IAO (Information Artifact Ontology)	244
IDO (Infectious Disease Ontology)	549
OBI (Ontology for Biomedical Investigation)	3,804
CL (Cell Type Ontology)	4,401
VO (Vaccine Ontology)	5,348
OAE (Ontology of Adverse Events)	2,558
NCBITaxon (NCBI organismal classification)	847,760
ERO (Eagle-i Reagant Ontology)	3,541
Cell Line Ontology (CLO)	38,689

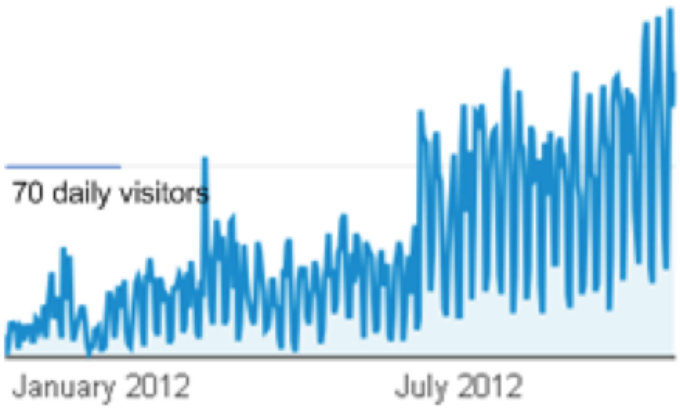


Figure 7.5: Records of Ontobee daily web page visitors according to Google Analytics.

	BioPortal	OLS	Manchester Browser	AmiGO	LLD	Bio2RDF	DBPedia	Ontobee
U1. All OBO,predictable	✓	✓	✓	n/a	✗	✗	n/a	✓
U2. Useful information	partial	✓	✓	✓	partial	✗	partial	✓
U3. Denotation clear?	✗	n/a	✓	✓	✗	✗	✗	✓
U4. IRIs citable	✗	n/a	✗	✗	✗	✗	✓	partial
U5. Accurate RDF	✗	n/a	n/a	✗	✗	✗	n/a	✓
U6. Both ontology & term views	✓	✓	✓	n/a	✗	✓	n/a	✓
U7. Readable OWL	✗	n/a	✓	✗	✗	✗	✗	✓
U8. Customized views	✗	✗	✗	✗	✗	✗	✗	✗
U9. Navigation tools	✓	✓	✓	✓	partial	partial	partial	✓
E1. Use of standards	✗	✗	partial	partial	partial	partial	partial	✓
E2. Predictable RDF assertions	✗	n/a	n/a	✗	✓	✗	✓	✓
E3. Re usable HTML	partial	✗	✗	✗	✗	✗	✗	✗
E4. Search visibility	✓	✗	✗	✓	✗	✗	✓	✗
E5. Scalability	✓	✓	✗	✓	✓	✗	✓	✓
E6. Transparency	✗	✗	✗	✗	✗	✗	✗	✓
Open source	✓	✓	✓	✓	✗	✓	✓	✓

Figure 7.6: Comparison of Ontobee features vs. other tools

7.3.6 Evaluation

Figure 7.6 compares the features of Ontobee and other tools. Ontobee has been designed and implemented with the requirements listed in Section 7.1.1 in mind. Ontobee provides a service with predictable behavior across the whole body of OBO library ontologies (U1). Although the instruction does not strictly require redirection to RDF/XML about the entity, this has been common practice. While less ambiguity in the specifications would be desirable, narrowing the range-14 approach by always returning RDF/XML about the entity is preferable. In Ontobee, useful information has been displayed (U2), and the ontology terms IRIs are clearly identified (U3). The Term IRI was bolded in Ontobee to suggest the usefulness of this term IRI for copy/paste and citation (U4). Care is taken to ensure that the RDF output is accurate and inclusive in terms of relation and axiom specifications compared to the source ontology (U5). In Ontobee, the term IRI is always delivered to the latest version. While there has been a tension between version used and unified view, it is possible to use the SPARQL endpoint and SPARQL queries to solve this issue. Ontobee presents well both ontology-centered and term-centered views through hierarchical tree visualization (U6). Readable OWL expressions are displayed in Ontobee RDF output (U7). Ontobee has not allowed customized views yet (U8), which will be considered for future Ontobee development (note: see more detail in the later Future direction part). Ontobee also provides ways to aid navigation to ontology terms of interest (U9). More methods can be developed to improve the implementation of this requirement, for example, providing links to graphic in OLS and links to where the term is used in GO.

Ontobee also follows the requirements on the engineering side. Ontobee adheres to the specifications of RDF, OWL, SPARQL, and XSLT (E1). It provides predictable access to RDF/XML assertions by accurately following the original ontology definitions (E2). Ontobee is able to generate HTML visualized and reused by other applications (E3). Allowing the customization of the HTML code will make Ontobee more powerful in this regard. Ontobee has not demonstrated good visibility in search results of popular search engines (e.g., Google) (E4). This is due to the lack of ability in indexing RDF search results in these search engines. Different approaches are being evaluated to solve this issue. Ontobee shows a comfortable scalability so far as explained in Section 7.3.4 (E5). Ontobee's performance will be monitored given with a possible significant increase of users in the future. Ontobee provides all SPARQL code for each ontology term IRI display supporting transparency and education (E6).

7.3.7 Future work

Currently Ontobee uses SPARQL 1.0 with Virtuoso extensions. Further development includes using SPARQL 1.1 and no extensions. More Ontobee features are expected to be developed. For example, the HTML rendering in Ontobee is currently not fully styled. Use of Cascading Style Sheets (CSS) would make it fully customizable so that individual ontologies could supply customized CSS. It will also be allowed to have a community-specific view choice so that projects like Neurolex [214] doesn't have to make new ids to have new pages. The RDF output content, i.e., which triples are included or excluded, can be improved with suggestions from Ontobee users and developers. Besides the current RDF contents for each ontology term, there are other alternatives to explore, for example, using modularity algorithms [174, 175] to construct a self-contained ontology that includes the term. Addition of other content types is planned, such as using foaf:depiction from the Friend of a Friend (FOAF) project [215] to show images. A Wikipedia setting may also be attached for users to comment, provide feedback, point to trackers and other resources.

Chapter 8

Representing pharmacovigilance data

8.1 Introduction

In this chapter, I describe how I developed the AERO, with the hypothesis that a standard and logically formalized representation of the Brighton Collaboration case definitions would enhance data quality and allow for automatic processing of adverse events reports. The AERO was used to encode the anaphylaxis case definition, the most complex Brighton guideline, which had previously been used in feasibility studies [15]. Several challenges that arose during implementation are detailed, such as definition of core terms (e.g., "adverse event"), relation between adverse events and the underlying biological entities (i.e., how does a finding of erythema relate to the physical manifestation erythema) and how to represent the assessment of an adverse event according to the Brighton guideline in a rigorous ontological framework.

8.2 Rationale for AERO and development practice

A formal and logical description of vaccine adverse events would allow their automated processing. For example, currently, software tools must deal with variation in the names of symptoms. A tool based on an ontology could present only relevant items and their definitions in a checklist, making it both easier to enter and validate data at reporting time. As detailed in 2.2.3, the availability of the AERO, an ontology representing the Brighton guidelines, in addition to the existing human readable format, would increase accuracy and quality of reporting. This, in turn, would facilitate further automated analyses of clinical data, potentially allowing detection of adverse events in a large population at a fraction of the time and cost currently incurred.

When developing AERO, care was taken to reuse, when possible, work done in the context of other efforts. Reusing terms from other resources allowed us to rely on knowledge of domain experts who curated them and to dedicate more work time for terms that need to be created de novo. When only few relevant terms were identified in an external ontology, these were imported using

the Minimum Information to Reference an External Ontology Term (MIREOT) guideline [30]. For example, in order to define vaccine adverse events, the VO [119] term *vaccination* [216], defined as "administering substance in vivo that involves in adding vaccine into a host (e.g., human, mouse) in vivo with the intent to invoke a protective immune response" is imported. That definition, in turn, uses the term *administering substance in vivo* [217] that OBI [83] defines. Similarly, the OGMS [132] has terms for pathological entities, diseases and diagnosis which AERO also uses as building blocks. In other cases, external ontologies have been imported as a whole: (i) the RO [78] contains a set of common relations, (ii) the IAO [125] deals with information entities and metadata, and (iii) the BFO is used as upper-level ontology. Finally, AERO is a driving effort for the Ontology of Medically Relevant Entities (OMRE) [218], to which it submits all signs and symptoms definitions, as those are not specific to AERO but rather intended to be used by other efforts. These resources are commonly used by the OBO Foundry [55] ontologies, of which AERO aims to be a part. Reusing terms from OBO Foundry Ontologies, where applicable, also improves the ability to interoperate with other resources that also use ontologies developed within the Foundry framework.

8.3 Guideline representation and evaluation in AERO

8.3.1 Adverse event class

Consider the following cases in which the clinician wishes to report adverse events:

- sensorineural deafness reported after measles, mumps, and rubella vaccination. This disturbance of the cochlea or auditory nerve results in hearing impairment, often loss of ability to hear high frequencies [219],
- infection such as in the case of leflunomide in treatment of arthritis [220],
- any of the dermatological adverse events observed in patients treated with etanercept [221],
- headaches reported following use of proton pump inhibitors such as lansoprazole [222],
- rashes, extremely common for example at the injection site.

These cases indicate that the type of an adverse event can be either of BFO's upper level classes - occurrent or continuant. OGMS currently defines *sign* as "A quality of a patient, a material entity that is part of a patient, or a processual entity that a patient participates in, any one of which is

observed in a physical examination and is deemed by the clinician to be of clinical significance." and *symptom* as "A quality of a patient that is observed by the patient or a processual entity experienced by the patient, either of which is hypothesized by the patient to be a realization of a disease.". Those classes are sibling of the *bfo:continuant* and *bfo:occurrent* classes, directly asserted under *bfo:entity*. Adverse events clearly match those definitions: they can be quality of the patient (for example, pallor or cyanosis), a material entity part of the patient (e.g., rash), or a processual entity that parts of a patient participate in (e.g., seizure).

Following this, *aero:adverse event* is logically defined as the union of *aero:adverse event process* and *aero:disorder resulting from an adverse event process* (i.e., the adverse event continuant described above). An *aero:adverse event process* is "a processual entity occurring in a pre determined time frame following administration of a coumpound or usage of a device"; this can be logically translated as (using the Manchester OWL syntax [110]):

```
Class: 'adverse event process'
    EquivalentTo:
        processual_entity
         and (preceded_by some
            ('adding a material entity into a target'
             or 'administering substance in vivo'))
```

where the classes *adding a material entity into a target* and *administering substance in vivo* are imported from the OBI [161]. The AERO definition of *adverse event process* is meant to be inclusive, and cover cases such as those described by the Manufacturer and User Facility Device Experience (MAUDE); for example the case of a patient fitted with bioprosthetic heart valves who dies within the following 4 months[22]. It is also worth noting that this definition of adverse event does not imply causation between the sign observed and the compound administration/device utilization, but is rather based on temporal association.

The adverse event continuant hierarchy was built under the *ogms:disorder* class (Figure 8.1), which is defined as "A material entity which is clinically abnormal and part of an extended organism. Disorders are the physical basis of disease." To avoid any language ambiguity by associating the terms event and continuant in the label of the class *adverse event continuant*, it was renamed *disorder resulting from an adverse event process*. As a general way of overcoming the potential issue between terms in use by clinicians and ontological usage in the context of the OBO Foundry, in

[22] http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMAUDE/detail.cfm?mdrfoi__id=1942591

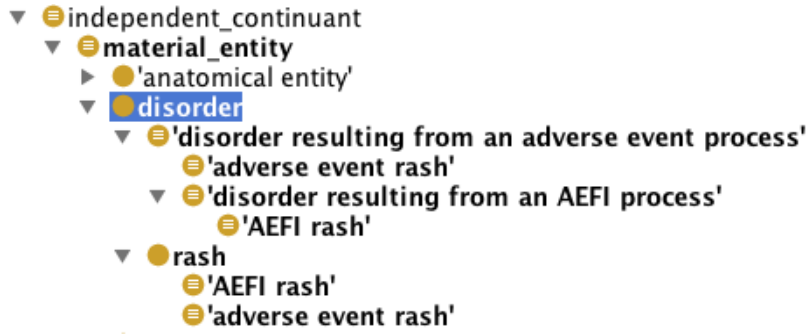


Figure 8.1: The disorder hierarchy as built in AERO, under the ogms:disorder class. The class *adverse event rash* is logically defined as the intersection of *disorder resulting from an adverse event process* and *rash*.

which it may be confusing to associate the word "event" to a hierarchical position under continuant, the *OBO Foundry unique label* IAO annotation property (`http://purl.obolibrary.org/obo/IAO_0000589`) is used. Classes such as *adverse event rash* (`EquivalentTo: disorder resulting from an adverse event process and rash`) will therefore have an OBO Foundry unique label annotation with value "rash resulting from an adverse event process".

8.3.2 Application of the guidelines

As discussed above, AERO is developed with extensive use of existing OBO resources. In order to present how I represent and compute with guidelines some orientation is first needed. Figure 8.2 depicts the representation of a patient examination, a typical way in which a set of findings is collected in post-licensing signal detection work. The process representation is from OBI. A patient examination is a planned process with (at least) three participants - the patient being examined, the clinician doing the examination, and the collection of findings created as a result. The class clinical finding is of information entities that are about medically relevant entities - material entities, qualities, processes, dispositions that are typically localized in an anatomical system or region. Medically relevant entities are to be considered a generalization of symptoms or conditions and are directly related to the patient or part of the patient. For example the entity *omre:low blood pressure* [223] is localized to the cardiovascular system. Clinical findings relate to the medically relevant entities and to the body systems using subproperties of *iao:is about* [224], a general relation between information and things in the world.

The collection of findings produced in the examination is an exam report. However generally

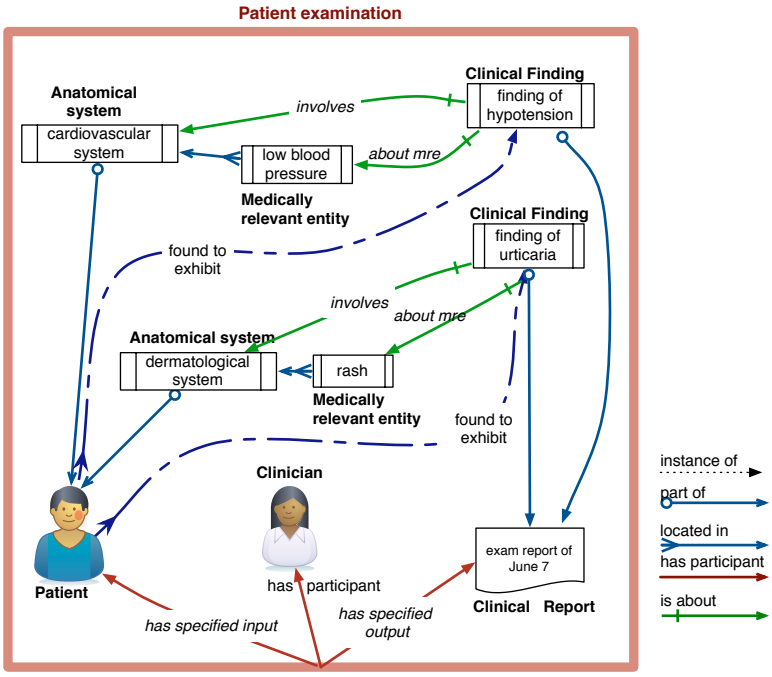


Figure 8.2: Entities represented in patient examination and recording of findings. During an *obi:planned process*(red surrounding box) a clinician examines a patient - the specified input- and produces a report which is a set of *ogms:clinical findings* - the specified output. Each finding *iao:is about* a medically relevant entity, mre (here a rash or low blood pressure) as well as the anatomical system or part proximate (here the skin or cardiovascular system). The report is a set of findings, each related to the report by the *aero:has component* relation.

speaking a distinction between reports, findings, diagnoses is not made. Each can have compositional structure, with parts related using *aero:has component* [225] and information about a patient that involves observation and judgment. A convenience relation *aero:found to exhibit* [226] is defined that relates a patient to findings about them. It is common to say that the patient has some finding but there is no essential relation from patient to finding. On the other hand the finding is dependent on the patient. In signal detection it is the exam report or a derivative of it that is the primary input to analysis.

8.3.3 Guidelines

Although there are a variety of kinds of clinical guidelines, the focus of the AERO is guidelines that are diagnostic in the sense that they provide, essentially, a recipe for taking a set of findings in the adverse event report and determining whether some specific medically relevant entity is implied

to exist. In the case of the Brighton guidelines the assessment also quantifies how certain one should be about whether the entity exists, by defining for example Level 1, 2 and 3 of certainty for the adverse events. A recipe is represented as an information entity, an *iao:directive information entity* [227]. The recipe and the Brighton case definition are related to the process of assessment by a composition of relations defined in IAO and BFO. As an information entity it is the sort of thing that can have many "copies". Each copy is represented as connected to the case definition using the *iao:is concretization of* relation. Such directive information entity is meant to be acted on, to be a representation of a plan - like a recipe. Plans (however they happen to be embodied) are represented using *bfo:realizable entity* [228], which connects the plan to a process in which the plan is carried out. The relation between the *bfo:realizable entity* and the process, should it occur, is called *is realized by*.

The current implementation accomplishes this classification by defining classes that correspond to the criteria by which each of the possibilities is determined. For example Brighton gives a set of conditions which, if obtained, provide the strongest evidence that a case of anaphylaxis has occurred *aero:level 1 of certainty of anaphylaxis according to Brighton* [229]. *aero:level 1 of certainty of anaphylaxis according to Brighton* is given a complete logical definition which is the expression encoding the criteria depicted in the lower middle of the figure. If the report has a set of finding components which together satisfy this class, then the report is classified as aero:level 1 of certainty of anaphylaxis according to Brighton.

The main classes used in the representation of guidelines are:

1. The *ogms:clinical finding* class [230]. A clinical finding is defined as "A representation that is either the output of a clinical history taking or a physical examination or an image finding, or some combination thereof." It does take into account historical information such as gathered from the patient's medical records, as well as results of assays such as blood tests or observations made about the patient by the physician. Clinical findings can themselves be diagnoses, allowing nesting of criteria as shown below in the case of uncompensated shock. A new relation, *found to exhibit*, has been created to link the patient to the clinical findings, such as "patient found to exhibit some nausea finding". It can also be used to link anatomical entities, such as a heart, to associated findings, such as "malfunctioning heart valve", allowing for diagnosis at multiple levels of granularity. In AERO, diagnosis are types of findings: it is often the case that the output of a diagnostic process is used to support further reasoning, such as when a physician establishes a diagnosis or respiratory distress based on a difficulty

breathing finding in a first step, and then relies on that respiratory distress diagnosis to infer, in conjunction with other findings, a diagnosis of anaphylaxis.

2. The classes of medically relevant entities from the OMRE. Those classes are of type pathological entities or formation as defined by the OGMS, as well as some processes. Signs and symptoms are separated from the assessment made of them to be able to consider them significant or not according to the specific guideline being used. For example, depending on the guideline considered, an increase in temperature will be considered a fever only if the temperature is above 37.8 °C (for example in older adult residents [231], or 38.3 °C [232] for neutropenic patients.

3. The anatomical entities which exhibits those findings. For example, *chest tightness finding* [233] involves the respiratory system, while a *measured hypotension finding* [234] involves the cardiovascular system. AERO doesn't define anatomical entities; rather they are imported from Uberon [235].

8.3.4 Anaphylaxis representation

In the AERO, the Brighton case definition for the anaphylaxis level 1 of certainty is modeled as an equivalent class as shown in Figure 8.3. It *has component* the different findings, grouped in sets according to their importance in the establishment of the diagnosis. For example, the major cardiovascular criteria set for anaphylaxis according to Brighton is the disjoint union of a clinical diagnosis of uncompensated shock and a measured hypotension finding. A clinical diagnosis of uncompensated shock is a clinical finding, but also a diagnosis established based on the presence of 3 or more uncompensated shock signs, but at most one of each type, as shown in the Manchester syntax [110]:

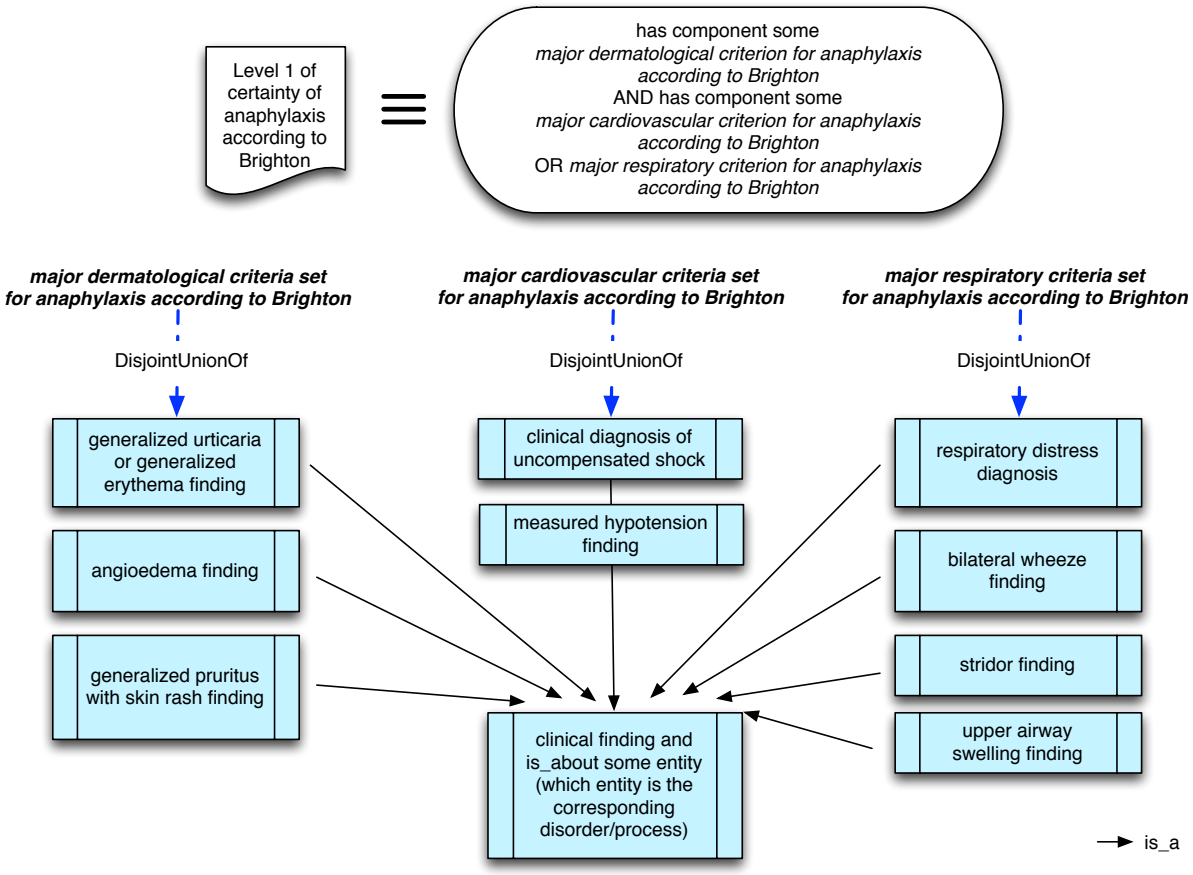


Figure 8.3: Details of the implementation of the level 1 of anaphylaxis according to Brighton. Sets of criteria are modeled as disjoint union classes, representing each of the findings that should be assessed by the physician.

```
Class: 'clinical diagnosis of uncompensated shock'
    EquivalentTo:
        ('has component' min 3 'uncompensated shock sign finding')
         and ('has component' max 1 'tachycardia finding')
         and ('has component' max 1 'capillary refill time > 3s finding')
         and ('has component' max 1 'reduced central pulse volume finding')
         and ('has component' max 1 'decreased level of consciousness
                        or loss of consciousness finding')

    SubClassOf:
        'clinical finding'
```

8.4 The has component relation

A relation, *has component* [225], was defined to relate clinical findings with the signs and symptoms that compose them. *has component* is a sub property of *has part* [236], which could not be used due to limitation on the use of non simple properties and cardinality restrictions[23]. Additionally, using *has part* didn't seem accurate; the clinical diagnosis of uncompensated shock doesn't have part a tachycardia finding, rather the finding is a component of the diagnosis.

8.5 The WHO severe malaria guideline representation

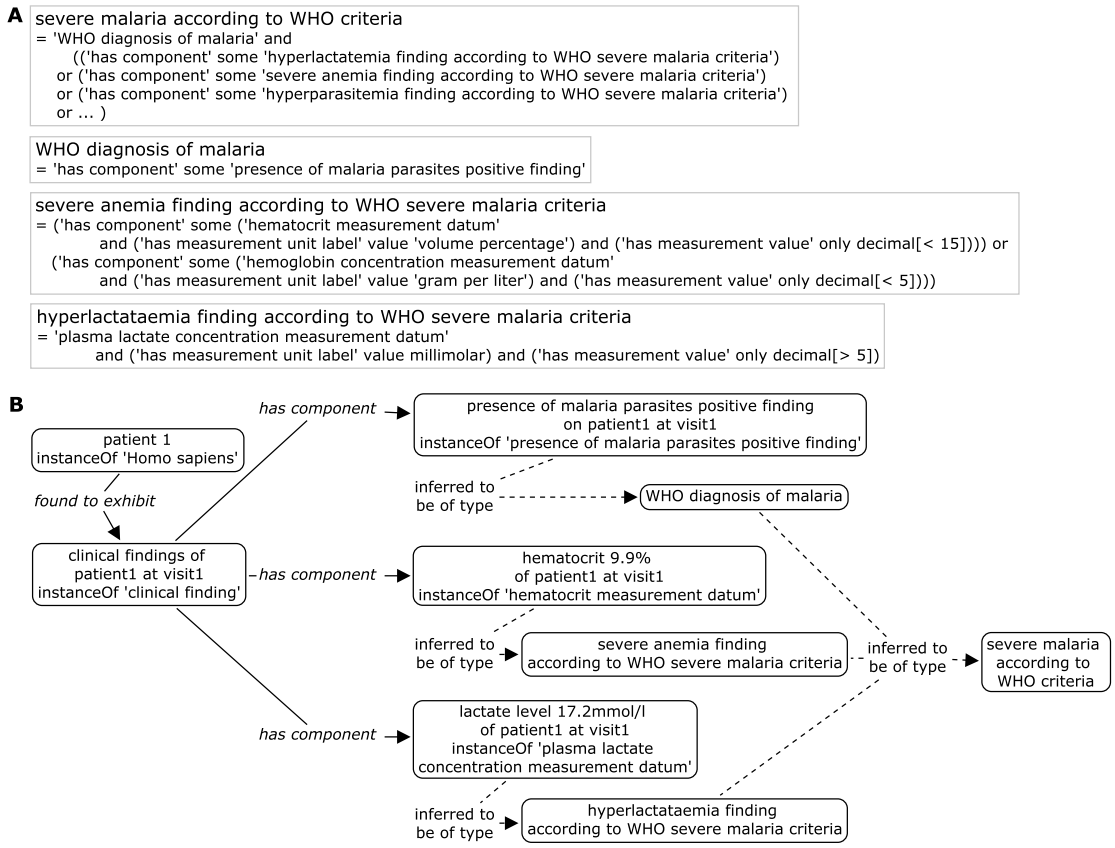


Figure 8.4: Implementation of the WHO severe malaria guideline. (A) Details representation of WHO severe malaria criteria as union of various criteria specified in the WHO guideline. (B) Example of classification: a diagnosis of severe malaria is inferred for patient1 based on laboratory data according to the WHO guideline.

[23] `http://www.w3.org/TR/2009/REC-owl2-syntax-20091027/#Global_Restrictions_on_Axioms_in_OWL_2_DL`

The WHO divides malaria into two categories, severe malaria and mild (or uncomplicated) malaria. Severe malaria is a life-threatening form of the disease requiring immediate hospital care and therefore correct classification of malaria is critical for appropriate patient treatment. The WHO specifies a list of criteria for severe malaria diagnosis [237]. The criteria include severe anemia, hyperparasitemia, hyperlactatemia, hypoglycemia, and over ten other different signs or symptoms. Severe malaria is diagnosed when any of the criteria are present. Otherwise, the diagnosis is considered to be mild malaria. Most of the symptom/signs are determined through laboratory tests and specified in the WHO guideline. For example, severe anemia is determined according to laboratory (or assay) results, hematocrit $< 15\%$ or hemoglobin < 5 g/dL and plasma lactate level greater than 5 mmol/l means hyperlactatemia [237].

The WHO severe malaria guideline is not as complex as the Brighton guideline as it does not need to relate symptoms and signs to specific anatomical systems. The severe malaria guideline does define symptoms and signs assessment based on laboratory measurement data in keeping with the approach described in the "Guideline representation in AERO" section but without a detailed implementation component. The *iao:scalar measurement datum* class [238] is used to represent measurement data to facilitate the diagnosis process. A scalar measurement datum is defined as "a measurement datum that is composed of two parts, numerals and a unit label". For example, hematocrit 17% can be logically represented as:

```
'has measurement unit label' 'volume percentage'
'has measurement value' ''17''^^decimal
   instance of 'hematocrit measurement datum' subClassOf 'scalar measurement datum'
```

The diagnosis pipeline for severe malaria is similar to the assessment of anaphylaxis level 1 according to the Brighton guideline. Applying the AERO developed pattern, severe malaria is modeled as the union of different criteria specified by the WHO. Formal and logical representation of severe malaria diagnosis and some related criteria using Manchester syntax is shown in the top part of Figure 8.4. It was tested by laboratory results and clinician's diagnosis published by Krupka, *et al.* [239].

8.6 Results

The pattern developed in AERO allows for automated classification of the patients based on a set of signs and symptoms they present, and the associated clinical findings assessed by their physician in

compliance with a selected guideline, as shown in Figure 8.2. Signs and symptoms are assessed by the physician during a patient examination, and the corresponding findings are of type generalized urticaria finding and measured hypotension finding respectively. These two clinical findings can then be inferred to be of type major cardiovascular criterion for anaphylaxis according to Brighton and major dermatological criterion for anaphylaxis according to Brighton. A diagnosis of level 1 of anaphylaxis is reached as they match the Brighton case definition for the components required.

Krupka, *et al.* [239] provided selected clinical findings and laboratory data of five patients associated with different malaria status. Using the WHO guideline, those patients were manually diagnosed as severe malaria and four of them with severe anemia before treatment. The automatic diagnostic classification results obtained from the implementation shown in section 8.5 are consistent with the manual assessment. The bottom part of Figure 8.4 shows detailed implementation of selected laboratory results associated with patient1 at the first visit. Based on clinical findings, the patient is classified as severe malaria according to the WHO criteria.

8.7 Discussion

It is critical in health care in general, and in analysis of adverse event in particular to be able to store medical data as well as the guideline that was used to assess it. Gagnon *et al.* [240] demonstrate that depending on the guideline considered, the number of anaphylaxis cases after injection of the adjuvanted H1N1 pandemic vaccine varies. The National Institute of Allergy and Infectious Diseases/Food Allergy and Anaphylaxis Network (NIAID/FAAN) considers that reduced blood pressure is enough to diagnose anaphylaxis after exposure to allergens [241], while two or more organ systems need to be involved as per Brighton. During vaccination, decrease of blood pressure is frequently caused by fear of the syringe or the vaccine, and may lead to false positives when diagnosed with the NIAID/FAAN guideline.

Knowing which guideline was used for diagnosis establishment is therefore important to be able to weigh cases as more or less important depending on their evidence and supporting or not detection of a safety signal and further actions by health authorities. An additional possible contribution is to allow for various versions of the same guidelines to be encoded. Different changes, such as scientific research progress, may warrant guidelines update [242], and it needs to be able to at a minimum accommodate their co-existence. Ideally, they could be partly reconciled, and facilitate migration from data encoded in the previous version to the newer one.

8.8 Conclusion

These results demonstrate that the pattern defined in AERO is applicable to the automated classification of AEFI according to the Brighton guidelines. It can be implemented in other applications, such as automatic malaria classification based on the WHO severe malaria guideline. The latter illustrates the potential to generalize the AERO diagnosis guideline pattern to formal and logical description of various diagnosis guidelines and facilitate automated disease diagnosis and validation. A standard representation of diagnosis criteria and clinical guidelines allows one to unambiguously refer to a set of carefully defined signs and symptoms at the time of data entry, as well as to choose an overall diagnosis that retains provenance links to its source, definition, and associated signs and symptoms. Such diagnosis is formally expressed, making it amenable to further querying for statistical analysis and other applications and supports query at different levels of specificity. Finally, cases encoded according to different guidelines may be reconciled; for example, based on their respective definitions, all cases of anaphylaxis according to the Brighton guidelines are also cases of anaphylaxis as per the NIAID/FAAN guideline (while the reverse is not true).

Chapter 9

Automated adverse events classification

9.1 Introduction

In this chapter, I apply the pattern developed in AERO, and described in Chapter 8, to large report collections from current reporting systems to allow those reports to be classified according to the Brighton criteria. This, in turn, will help identify potential cases on which human review should be focused and decrease cost and time by reducing manual evaluation.

Currently, efficient analysis of adverse event reports is a time-consuming task, requiring qualified medical personnel. For example, a team of 12 medical officers worked for over three-months to review 6,000 post-H1N1 vaccination reports for positive cases, only a fraction of the total number of reports received [243]. Ideally, enabling automatic case classification from specialized reporting systems such as the VAERS [244] used in the United States and the CAEFISS implemented in Canada would allow analysts to confirm or discard diagnoses made by physicians and identify additional probable cases for further investigation. However both those datasets are imperfect. While the Brighton guidelines have been adopted as standard by PHAC, their usage in practice is scarce. They are not implemented in the reporting pipeline, but for a partial implementation in the form of check boxes in a PDF form [245]. In practice, the free text part of the reports is manually annotated, and part of the reports is then reviewed by medical experts. Additionally, both VAERS and CAEFISS currently rely on MedDRA to encode adverse events data. This section describes how, using a mapping to convert MedDRA codes to AERO annotations, I was able to process the existing MedDRA annotations on the data and infer if a Brighton criteria has been met or not, as shown on Figure 9.1.

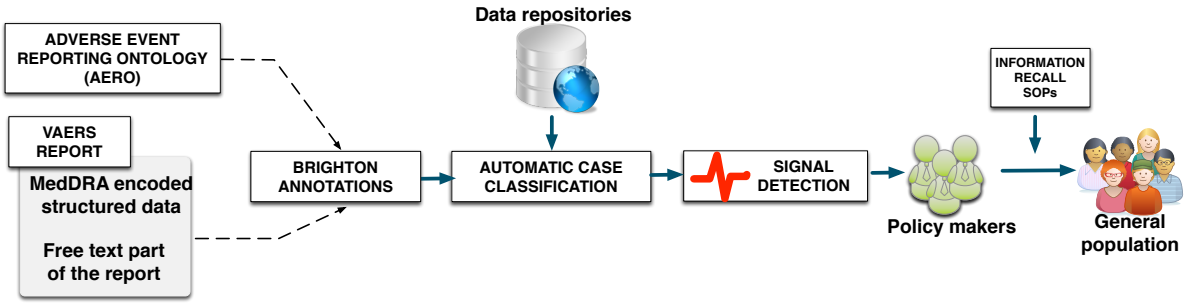


Figure 9.1: Automatic case classification according to the Brighton criteria. Classified case reports allow for signal detection and policy makers information, impacting public health.

9.2 AERO ontology

In Chapter 8 the development of the AERO was described. Here I present how it is being used in practice to enable automation of adverse events classification, by assessing whether they correspond to the Brighton case definition criteria.

9.2.1 Assessment pipeline

Figure 9.2 shows how the various entities are related in AERO to form a diagnosis pipeline for the assessment of anaphylaxis according to the Brighton guideline. The patient examination by the physician results in a set of clinical findings that are part of a report, upper left. The report findings are input to a process of diagnosis which uses the case definition. The case definition is *concretized as* the plan to use the guidelines in a process of diagnosis, and that this process *realizes the plan* (in figure as *manifests as*). The case definition includes different criteria concerning the findings, each of which, when satisfied, yields some assessment of the certainty of anaphylaxis being present. For example, the lower middle stack represents the criteria for diagnosing a *level one of certainty according to the Brighton anaphylaxis guideline*. When findings in the report together satisfy these criteria, the output of the diagnostic process is determination of the level 1 of diagnostic certainty of anaphylaxis according to Brighton.

Figure 9.3 gives two extracts from the class hierarchy related to terms in the figure. It reads: Every 'level 1 of diagnostic certainty of anaphylaxis according to Brighton' is a 'Brighton diagnosis of anaphylaxis as an AEFI', which is in turn is a 'Brighton diagnosis', itself a 'clinical finding'. Every 'Brighton case definition of anaphylaxis as an AEFI' is a 'Brighton case definition', which in turn is a 'diagnosis guideline'.

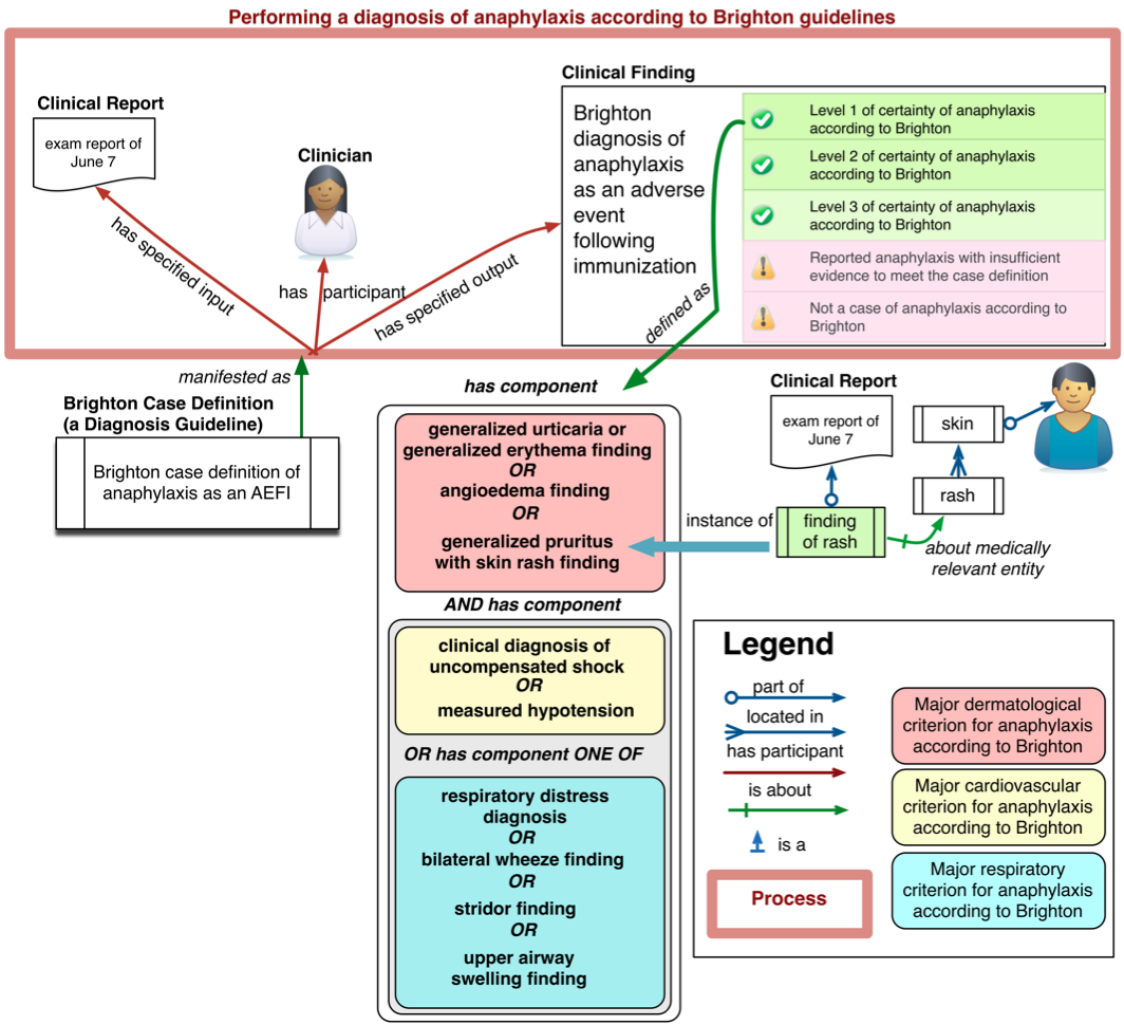


Figure 9.2: The elements of an assessment of anaphylaxis according to Brighton as implemented in AERO. Performing a diagnosis involves assessing a number of criteria each (e.g., lower middle box) implemented as a class expression that classifies a set of findings. The diagnosis of Level 1 of certainty of anaphylaxis is made by the clinician if the written criteria apply, and by the OWL implementation if the class expression subsumes the set of findings shown in illustration as a *Clinical Report*.

9.3 VAERS dataset

The VAERS [26] is a post-market passive surveillance system, under joint authority from the Center for Disease Control and Prevention (CDC) and US Food and Drug Administration (FDA). It provides self-reporting tools for individuals and health practitioners, and its datasets are publicly available. VAERS reports are semi-structured. A free text field contains the report notes, and

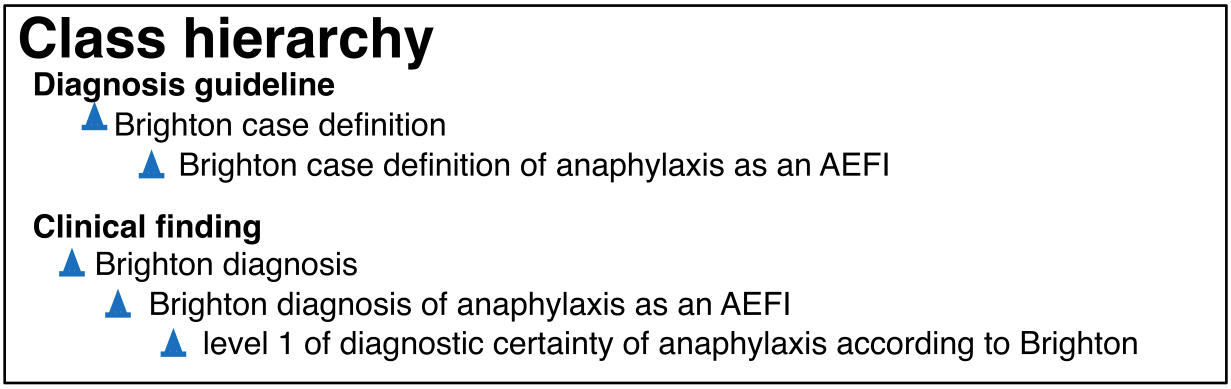


Figure 9.3: Class hierarchy excerpt in the AERO. Every 'level 1 of diagnostic certainty of anaphylaxis according to Brighton' is a 'Brighton diagnosis of anaphylaxis as an AEFI', which is in turn is a 'Brighton diagnosis', itself a 'clinical finding'. Every 'Brighton case definition of anaphylaxis as an AEFI' is a 'Brighton case definition', which in turn is a 'diagnosis guideline'.

another field contains a list of MedDRA terms that correspond to the report.

The dataset described in [243], was obtained through a series of Freedom of Information Act (FOIA) requests. It consists of 6034 reports received between the end of 2009 through early 2010, all following H1N1 vaccination after the FDA was alerted of a possible anaphylaxis safety signal by the PHAC. However data surrounding the 100 confirmed anaphylaxis cases in the original report were unobtainable as they were deemed lost. All reports in this set were evaluated by specialists and so provide a gold standard for comparison. A series of FOIA requests were also used to obtain the dataset describing classification results on the same dataset using the ABC tool, the MedDRA Standardized MedDRA Queries (SMQs) as well as a custom information retrieval method [246]. However details of the original analysis approach necessary for reproducing the original results were not made available and I could only hypothesize the cause of results obtained that were not in concordance with the original publication.

To demonstrate that the AERO can be used to effectively encode a logical formalization of the Brighton guidelines, the output of classification using the ABC tool with the results of the classification using the ontology was compared.

9.4 Data loading and processing

To streamline the analysis process, Python was used to perform the following steps, semi-automatically:

1. Load the VAERS reports into MySQL. The VAERS data was provided as a set of Excel spreadsheets, and MedDRA is distributed as ASCII files and corresponding database schema. Both were loaded into a relational database for easier processing.

2. Apply the mapping ([246], Electronic Supplementary Material, Appendix 3) from the existing MedDRA annotations to the Brighton terms. Each MedDRA ID was mapped to the corresponding AERO ID, and a mapping table was created in the database.

3. Export the dataset into a series of RDF files and perform pre-processing. As working with the complete dataset in OWL was neither efficient nor necessary the data into smaller files was partitioned as follows.

4. Export the dataset into a series of RDF files and perform pre-processing. For each report (i.e., each VAERS ID), all information in that report was collected for RDF serialization

   Next MedDRA terms were mapped to assertions using AERO. Because the OWL representation required more information than was available in the reports choices had to be made before classification could proceed, specifically (1) setting some Brighton required values to true as they cannot be encoded in the current version of MedDRA (2) add negation to reports to simulate the closed world assumption made in the reports. These steps are both further explained below.

   Serialization was done using the FuXI framework [247], which provides a syntax for OWL [160] entities in Python that is more amenable to coding than RDF/XML.

5. Apply an OWL reasoner to classify reports. The reasoning step was performed with the HermiT reasoner [101], via the OWLAPI [150]. In series, each RDF file was loaded, the reasoner computed inferred axioms, including individual types assertions, and those axioms were recorded into another RDF file.

6. Load each of the original RDF and associated inferred axioms as well as AERO into a Sesame triplestore [106]. I found it was more user friendly to use Sesame's interface for querying.

9.5 Brighton classification results

I was able to successfully classify a subset of just over 6000 VAERS records in just over 2h on a Mac OS X laptop with a 2.4Ghz Intel Core i5 and 8GB of memory. The triplestore was then queried to

Table 9.1: Classification results. The first row are the results of running the ABC tool online, as described in [246]. The second row is the initial ontology-based classification, using the same rules and with the addition of the negation for information not present in the reports. The last row is the ontology-based classification without the addition of the negation. Level 1, 2 and 3 columns represent the existing Brighton classification categories. Level 2 updated and Level 3 updated represent the category as they should have been encoded based on communication with the Brighton collaboration.

	Positive cases					Negative cases		
	Level 1	Level 2	Level 2 updated	Level 3	Level 3 updated	Insufficient evidence	Not a case	No evidence
ABC tool	101	221	N/A	7	N/A	488	2844	2373
Ontology with negation	98	223	223	8	8	3	3078	2622
Ontology without negation	98	178	223	4	8	3	3078	2622

retrieve reports in each of the Brighton case definition categories; results are shown in Table 9.1.

However, three issues were identified, either with the annotation standard being used (such as MedDRA), the quality/availability of the information in the reporting systems (such as VAERS) and interpreting the guideline (such as Brighton case definitions).

First, there are critical limits to the temporality representation in MedDRA. Temporality information is needed for causality assessment. It is a necessary (though not sufficient) condition that the temporal association be consistent with the vaccination. Temporality data is also needed for diagnosis determination (which is of interest for the classification) to represent dynamic disease conditions, such as onset, progression (rapid, chronic?) and relapsing. In the specific case of anaphylaxis, there are no MedDRA terms allowing encoding of 'sudden onset' and 'rapid progression' which are necessary conditions to reach any positive level in the Brighton classification of Anaphylaxis. The strict application of the Brighton guidelines to the VAERS dataset as-is would result in a value 'don't know' for those criteria, and consequently classify all reports as negative (insufficient evidence/not a case).

Second, there are no distinctions between unknown/missing/non applicable information in the reporting systems. In the case of 'generalized pruritus without skin rash', when the report does not provide any information about 'skin rash', it is impossible to know whether that information is unknown (the physician did not check for presence/absence of skin rash), missing (the physician did check but the information was not recorded) or was negative and therefore not included in the report (the physician checked and did not see a skin rash, but the negative finding was not included in the report).

To remedy those two major issues, and for the purpose of research, the condition that 'Rapid progression' and 'sudden onset' criteria are not required for diagnosis was added to the VAERS dataset. Also the negation of those signs or symptoms that were not positively stated on each report was added.

For example, clinical findings of the report 369695 are defined as (shown using the Manchester syntax [110]):

```
Individual: 369695

    Types:

        'clinical finding',

        'has component' some 'generalized erythema finding',

        'has component' some 'generalized urticaria finding',

        'has component' some 'difficulty breathing finding'
```

does not reach a Brighton level of diagnosis certainty. However, with the addition of the restrictions

```
not ('has component' some 'bilateral wheeze finding'),

not ('has component' some 'stridor finding')
```

the condition for minor respiratory criteria is fulfilled ('difficulty breathing without wheeze or stridor') and the report is classified as Level 2 of certainty.

Third, when translating the Brighton guidelines into their logical form, different interpretations of the same human readable content were observed, and I conducted extensive discussion with the Brighton collaboration to clarify the formalization due to this ambiguity.

Upon realizing that the addition of negation to the dataset would be required (that I established was also the case in [246], though unpublished), further enquiries were made with the Brighton collaboration as to whether those negations were logically and clinically required or if the were added to allow human readers to distinguish between minor and major criteria. For example

'pruritus with or without skin rash', which is major or minor criterion respectively: 'pruritus' ought to be enough as minor criterion, there should be no need to require the presence of the 'no skin rash' (which is currently required in the ABC tool). Practically, this means that when considering a report annotated with 'Rapid progression of signs and symptoms', 'Sudden onset of signs and symptoms', 'Hypotension, measured', 'Pruritus, generalized': with the addition of 'Skin rash: Yes' it classifies as expected as Level 1. With the addition of 'Skin rash: No' it does classify as expected as Level 2. However, with the addition of 'Skin rash: Don't know' it classifies as 'insufficient level of evidence' - which is incorrect: even if it is unknown, there was either presence of skin rash or not, so this report should at a minimum classify as level 2 of diagnostic certainty. Another outcome of this work is that compound terms should be represented as association of individual terms. For example, 'capillary refill time of >3s without hypotension' should be encoded as 'Capillary refill time > 3 sec' and not 'Hypotension, measured'. There are currently 2 entries in the ABC tool: one can either select 'Capillary refill time > 3 sec' Yes and 'Hypotension, measured No' OR one can select 'Capillary refill time > 3 sec, no hypotension'. While the former behaves as expected when applied to an anaphylaxis report for which 'capillary refill time of >3s without hypotension' is a cardiovascular criterion, the latter doesn't allow for correct classification. Similarly, in the pruritus case above, 'generalized pruritus with skin rash' should be 'generalized pruritus' and 'skin rash'. This allows differentiating between a major dermatological criterion ('generalized pruritus with skin rash') and the corresponding minor dermatological criterion ('generalized pruritus' and not 'generalized pruritus without skin rash'). By systematically reviewing and applying this to other criteria, I was able to overcome the need for addition of negation in the dataset. This can be more or less complex depending on the number of such negated criteria in the original case definition. Also, there exist different human interpretations of the same guideline, often linked to ambiguity in the textual representation of the criteria. For example, the case definition of anaphylaxis (described in Table 2.2) states that a level 3 of diagnostic certainty is reached when the following are observed:

- ≥1 minor cardiovascular OR respiratory criterion AND
- ≥1 minor criteria from each of ≥ 2 different systems/categories

This was interpreted as (1 minor cardiovascular OR respiratory criterion) AND 2 minors from systems that are neither respiratory nor cardiovascular (dermatologic, gastrointestinal, laboratory systems) and so translated in the ABC tool. However it should have been read as "if there is

a minor cardiovascular criterion, then 2 other systems need to be involved, including respiratory, dermatologic, gastrointestinal and laboratory" (and vice versa for a respiratory criterion).

Following discussion of those results with the Brighton Collaboration, an updated version of Brighton guidelines was encoded and added in the AERO, in addition to the existing ones, to reflect those changes. Based on these changes I was able to reason over the dataset, without the addition of negation, and simultaneously compare the different cases, shown in table 9.1 under columns 'Level 2 updated' and 'Level 3 updated' (there were no modifications to the Level 1 or associated criteria). Using the updated logical translation of the Brighton guidelines, the intended results were achieved. In the row 'Ontology without negation', there are 223 cases for 'Level 2 updated' and 8 results for 'Level 3 updated'.

By comparison the original algorithm misses cases and detects only 178 cases for 'Level 2 updated' and 4 results for 'Level 3 updated' (20% and 50% missed respectively). Finally, a rather large difference was observed for the category 'Insufficient evidence'. Running the ABC tool as shown in [246], Botsis et al. found that 488 cases were classified as 'Insufficient evidence'. However, according to the Brighton guideline, the full label for this category is 'reported anaphylaxis with insufficient evidence', and is meant to identify cases for which there may have been misdiagnosis from the reporting physician, or not enough evidence according to the Brighton criteria to establish the anaphylaxis diagnosis. In the original dataset, only 12 reports were annotated with an 'anaphylaxis' MedDRA term (including anaphylaxis-like terms, e.g., anaphylactic reaction, anaphylactic shock,...). Out of those 12 reports, only 3 were lacking supporting evidence as shown in Table 9.1, column 'Insufficient evidence', rows 2 and 3. This results from the fact that the online ABC tool that was used for classification, provides a 'diagnosis confirmation' tool, which implies that the user wants to confirm an anaphylaxis diagnosis that they established. Consequently, those 488 cases were incorrectly categorized as 'reported anaphylaxis with insufficient evidence'.

9.6 Automated case screening

In the previous section the Brighton guidelines were translated into their logical representation, and applied the AERO to automate classification of vaccine adverse event reports from VAERS. As shown in Table 9.2, while the resulting specificity is very high (97%), the corresponding sensitivity is fairly low (57%).

This can however be easily understood remembering that the Brighton guidelines were never

Table 9.2: Comparison of different classification methods. * indicates that the result was taken from [246] (values for the testing set). In the Brighton Collaboration section, the ABC tool and ontology-based classification have similar outputs (the small difference in terms of sensitivity can be explained as Botsis *et al.* split their dataset into training and testing). In the SMQ section, the expanded SMQ yields better results in terms of sensitivity and specificity compared to the existing SMQ categories and the IR approach proposed in [246]. CI: confidence interval.

	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)
Brighton Collaboration			
ABC tool*	0.64 (0.52-0.75)	0.97 (0.96-0.98)	NA
Ontology Classification	0.57 (0.51-0.64)	0.97 (0.96-0.97)	0.77 (0.74-0.80)
IR approach*	0.86 (0.75-0.93)	0.7861 (0.76-0.80)	NA
SMQ			
SMQ categories (combined)*	0.54 (0.42-0.66)	0.97 (0.96-0.98)	NA
IR approach*	0.85 (0.73-0.92)	0.86 (0.84-0.87)	NA
Expanded SMQ	0.92 (0.89-0.95)	0.88 (0.87-0.89)	0.96 (0.95-0.97)

meant for screening, but instead are reporting and diagnosis confirmation guidelines. The guidelines themselves were designed to identify only portion of the cases (low sensitivity) but do so extremely accurately (high specificity). Sensitivity needs to be significantly increased for the purpose of automated identification of rare adverse events. To address the issue of detecting similarity between the diagnosis text and the adverse event reports, the well-established information retrieval technique of cosine similarity [248] was used. Each document (gold standard query or report) was decomposed into its corresponding vector of terms (e.g., 'skin rash', 'generalized pruritus'). The angle those vectors form can be used to measure the similarity between them: the cosine of the angle is 1.0 for identical vectors and 0.0 for orthogonal ones. Terms in the vectors were weighted using the term

frequency-inverse document frequency (tf-idf) scheme, which numerically translates the importance of each term in function of its frequency (tf) in a given document and its frequency in the global dataset (idf). This method can be used to compare the vector terms extracted from each adverse event report against the chosen gold standard, such as Brighton or MedDRA terms. In [246], the authors divide the whole dataset into training and testing subsets (details of which are unpublished), and use the training subset to identify which terms are correlated with the outcome, which they then use to classify reports in the testing set. This method leads to a 85% sensitivity and 86% specificity (Table 9.2, section SMQ, row IR approach).

Upon inspection of the MedDRA SMQ and MedDRA terms used to annotate the reports, I realized that some of them which should presumably be highly correlated with an anaphylaxis diagnosis (e.g., "hypersensitivity") were not included in the existing MedDRA SMQ and therefore not considered for diagnosis assessment. Therefore, rather than creating a bag of words *de novo* based on keyword extraction from a training set of reports, I chose to expand a known already widely implemented screening method, i.e., the MedDRA SMQs. To identify which terms statistically correlate significantly with the outcome, the 2273 different MedDRA terms were extracted, and, using the classified dataset, for each a contingency table was built and the associated χ^2 and p-value computed, as shown in Table 9.3.

An α level of significance at 0.05 (arbitrarily chosen) and at one degree of freedom corresponds to a χ^2 value of 3.841.

Table 9.3: Contingency table per MedDRA term

	MedDRA term x	Not MedDRA term x
Anaphylaxis	a	b
Not anaphylaxis	c	d

The 120 MedDRA terms above this threshold were selected (see Appendix G), to which the 77 terms from the existing MedDRA SMQ were added, and then duplicates removed. The remaining 168 MedDRA terms were used to perform the cosine similarity based classification: they form the gold standard vector against which each of the report vector will be compared against. I first performed the analysis using a 50/50 training/testing data split: half the dataset (training) was used to build the MedDRA contingency tables, and classification was performed on the second half of the data (testing). The cosine similarity values obtained for each report were used to build a

ROC, and the best threshold value was obtained using the shortest Euclidean distance between the curve and the top left corner as well as the Youden index. At the best cut-off point (r = 0.051) I obtained 92% sensitivity (86-96% at 95% CI) and 81% specificity (80-82% at 95% CI) in the testing set, AUC 0.93 (0.9-0.95 at 95% CI). I then classified the whole dataset, and as shown on Figure 9.4, this expanded MedDRA SMQ significantly improves sensitivity (92% against 85% in [246]) with slight increase in terms of specificity (88% against 86%). The Area Under the Curve (AUC) was also high (0.96) compared to 0.80 in Botsis *et al.*'s training set: using my approach the classifier correctly discriminates between a positive and negative outcome in almost 96% of the cases.

Full classification results are shown in Table 9.2.

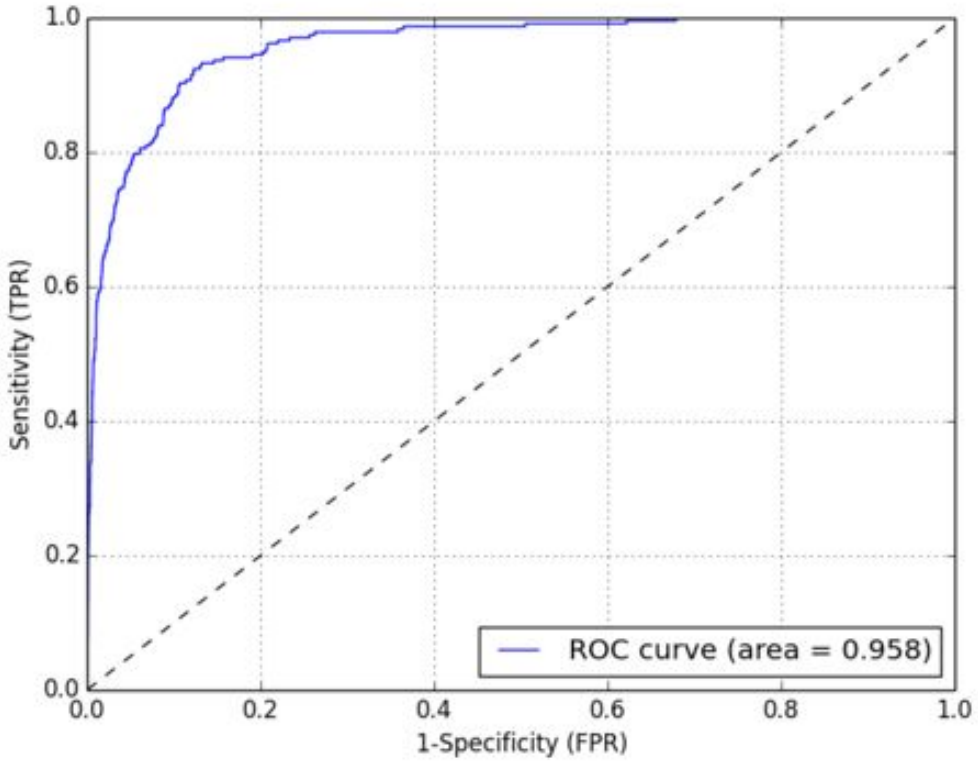


Figure 9.4: Cosine similarity ROC curve. ROC curve showing the sensitivity (True Positive Rate, TPR) vs. 1- Specificity (False Positive Rate, FPR) when measuring cosine similarity of the expanded MedDRA SMQ built from the existing SMQ and augmented with the terms identified as being significantly correlated with the outcome based on contingency tables. Statistics were computed using the R pROC package [249].

9.7 Discussion

These results indicate that using a logical formalization of existing guidelines helps identify missing elements in the reporting pipeline, as well as errors in the interpretation and application of the guidelines. Also the Brighton guidelines are not optimally suited for case identification in the currently existing reporting systems. Despite having an efficient, standardized and accurate ontological representation of the information, the guidelines were not designed for this purpose. By providing a suitable formalism and method, and encoding multiple versions of the Brighton guidelines, I demonstrated that the AERO can represent multiple guidelines, and allows for immediate comparison of classification across them. Additionally, this work suggests that relying only on the MedDRA encoded anaphylaxis (and associated synonyms such as 'anaphylactic reaction') in VAERS [250] may cause severe underestimation of the number of actual cases, as it was found that only 12 reports were reported as anaphylaxis in a dataset in which careful manual review identified 236 potentially positive cases. Finally, I demonstrated that automated adverse event screening can reach a very high sensitivity and specificity by building a specific bag of words (SMQ or guideline based) for each AEFI, on the best query terms I identified.

9.7.1 Using an OWL-based approach

Current state of the art for automated use of the Brighton case definitions is the ABC tool; however as shown above it is not suitable for automated classification. My approach not only addresses the limitations of the ABC tool, but also provides an open and extensible foundation which can be incorporated into future classification tools. Despite the Brighton guidelines not being optimally suited for the screening problem in the current context, there are multiple benefits in choosing to adopt a logical formalization of the surveillance guidelines considered, detailed below. Regarding the choice of the formalism, OWL is an accepted standard for knowledge representation, and comes with a large suite of tools allowing editing, storage and more importantly reasoning is supported by various softwares [60, 101, 103, 126, 150, 160]. This work demonstrates that even complex guidelines, such as the Brighton Anaphylaxis one, can be encoded using OWL2, and successfully lead to the desired inferences.

9.7.2 Limitations of the results

The main limitation of the results is that only the reports' annotations are analyzed. The ability to use Natural Language Processing (NLP) methods on the textual part would potentially allow further discrimination, and provide supporting evidence in decision making. Additionally, a mapping between MedDRA and Brighton was used for part of the classification pipeline. This mapping is subjective and may not be identical to the one another group would produce. Finally, while I could have worked towards increasing the sensitivity/specificity of the classification results using the AERO, I decided that this would change the purpose of the Brighton guidelines and was not desired. However, one could imagine that a 'Brighton screening guideline' could be created for that purpose.

9.7.3 Formalization of the case definition

Having a formal representation of the guideline, which could be distributed alongside a manuscript, would help both prevent misinterpretation (such as those observed as a result of not taking into consideration the underlying assumption that it performs diagnosis confirmation), and enable homogenized implementation in electronic systems of the chosen standard. Several studies [251, 252] rely on the number of adverse events detected in VAERS to hypothesize whether their rate is higher than expected with a certain vaccine. It is not currently possible to compare those studies, not even in cases in which they concern the same adverse event. For example, in [253], the authors define anaphylaxis in a less restrictive way than the Brighton criteria. In [254], yet another set of criteria is used, even though the two papers share authors. In [255], the authors acknowledge that different criteria were used for anaphylaxis identification, including the Brighton criteria, but conclude that they could not use the latter as this was not compatible with existing published reports. It is critical to ensure that not only reporters use standard for reporting, but also that medical officer know which standards were used, and be able to compare different ones. This is not only crucial for VAERS, but also, and more importantly, critical to reach the goal of having an international assessment of vaccine safety [256]. Finally, several projects have been recently concerned with addressing the need for reporting guidelines, such as the CARE guidelines [257], the PROSPER Consortium guidance document [258] or the integration of guidelines into asthma electronic record [259], the latter two specifically advocating for the use of taxonomies.

9.7.4 Time gain in signal detection

The approach I developed allows for earlier identification of a safety signal indicating a high level of adverse events related to vaccination, potentially preventing further adverse events. Figure 9.5 illustrates this time gain using the ontology-based method over the manual analysis. The VAERS dataset comprises just over 6000 reports which were collected over 2 months, and required 3 months for manual analysis by 12 medical officers [246]. By contrast, those 6000 reports can be analyzed almost instantaneously using the ontology based, automated approach - the only delay is due to the time needed to accumulate enough reports for analysis. As a result, in this case, the time gain would be at least a month during the flu season, which could translate in earlier detection of a safety signal, and subsequent forwarding of the information to relevant health authorities. Whether to automate the process of adverse event reports analysis is a health policy decision. It can be hypothesized that increase in cost and/or number of reports (for example as more provinces adopt an electronic reporting system) are two critical factors.

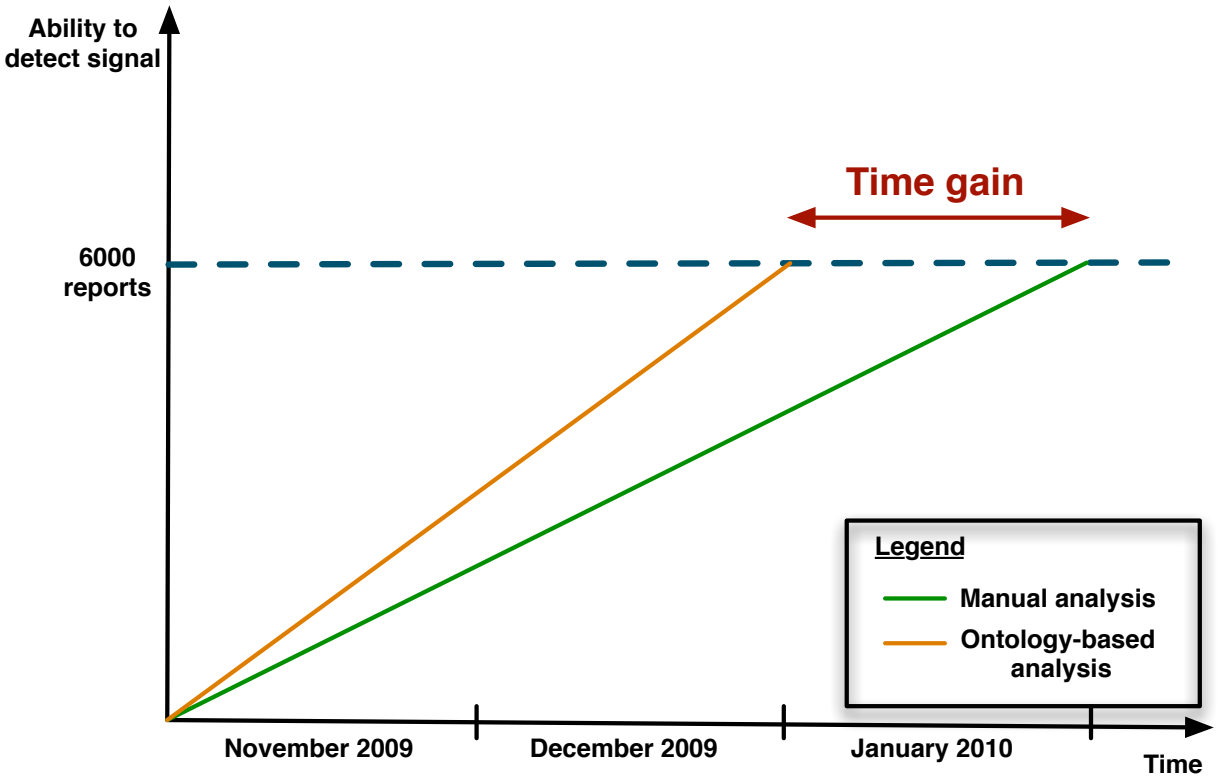


Figure 9.5: Time gain using the ontology-based method. As soon as the 6000 reports in the VAERS dataset are accumulated (2 months) they can be automatically analyzed, by contrast with the manual analysis which requires 3 months for 12 medical officers.

9.7.5 Use of the ontology for reporting

Another way to improve detection of adverse events is to standardize the reporting step. Currently, reports are centralized and then annotated with MedDRA terms by specialized coders. These individuals do not see the patient, and if deemed it necessary, they need to request more detailed medical reports after the fact. A tool that allows unambiguous and consistent reporting of the signs and symptoms they observe was provided to the person reporting the event, at data entry time, this information could be captured within the submitted report, and subsequently complex data-mining of the reports to classify them would not be needed. Using the ontology at data entry time would provide two distinct advantages: (1) the ontology provides textual definition for all the criteria terms and (2) the ontology can be used to enforce consistency checking at data entry time. Regarding (1), one of the requisite of my collaboration with PHAC was that the resource developed would be usable by human as well as machines. Not only were the logical axioms derived from the Brighton case definitions encoded, but also the human readable labels and textual definitions were added, most of those provided from [15]. Regarding (2), upon development of a data capture form capturing the Brighton criteria, the ontology can be used locally to check whether conditions for the diagnosis establishment are met. For example, when a physician reports 'anaphylaxis', the system could automatically ask for relevant signs and symptoms and store whether they have been observed or not. This would also help with respect to capturing whether the information that is not present in the report is missing or unknown.

9.7.6 Going forward: proposed implementation

As rare adverse events are considered, there is a need to ensure all possibly potential cases are retrieved, and to the best of my knowledge these results are the best obtained to date. I recommend a hybrid approach where both the SMQ information retrieval method and the AERO classification approach be used in parallel. The output of the high sensitivity classifier allows for extraction of a subset of the original dataset, even though there will be false positives (12.3%). Here, 5082 true negatives were rightfully discarded. If intersecting, the Brighton confirmed cases can be subtracted from this, allowing curators to focus on the remaining reports. Also, a fast screening method when data is being sent in would allow to automatically identify potentially positive cases, at which point a more detailed form (such as the Brighton-based reporting form from PHAC) can be immediately provided to the reporter.

9.8 Conclusion

By standardizing and improving the reporting process, the diagnosis confirmation was automated. By allowing medical experts to prioritize reports such a system can accelerate the identification of adverse reactions to vaccines and the response of regulatory agencies. Future reporting systems should provide a web-based interface (or a form in their electronic data capture systems) that reflects the criteria being used for case classification. This would help ensure that the information being captured is standardized and that potentially missing information can be immediately added by adding consistency checking tests. While this chapter provides way of improving standardization in passive, spontaneous reporting systems such as VAERS, other avenues can be explored to improve surveillance, such as promoting active systems [260]. At a minimum, providers of guidelines should recognize issues such as those described here, and commit to provide logical representations of their work. Based on our partnership and results, the Brighton Collaboration is moving towards providing such a representation for their case definitions.

Chapter 10

Conclusion and future directions

10.1 Summary

The first part of my thesis shows how I co-developed multiple ontological resources, focusing on the OBI in Chapter 3 and the VO in Chapter 4. Various use cases were presented, each exemplifying application in a different domain, demonstrating that ontologies allow for unambiguous and standardized representation of biomedical knowledge.

The second part of my thesis describes collaborative development in the context of the Semantic Web. Building large, interoperable ontological resources necessitated addressing some issues such as enabling rapid addition of similar terms following a pre-established pattern (QTT, Chapter 5), devising common policies and guidelines (OBO ID policy and IAO metadata in Section 6.2) and generally supporting reuse of existing ontologies to avoid duplication of efforts and multiplicity of URIs (MIREOT and OntoFox, described in Sections 6.3 and 6.4). Publication of those resources to improve their visibility and make them available via the Semantic Web was realized via the Ontobee, described in Chapter 7. Finally, the last part of my thesis shows that, relying on these efforts, adverse event classification in pharmacovigilance can be improved and automated through the use of ontologies. I built the AERO to encode pharmacovigilance guidelines and data (Chapter 8) and validated it against a manually curated dataset, demonstrating high specificity in Chapter 9.

10.2 Perspectives and future work

10.2.1 Coordinated maintenance of resources

Disappearance of online resources in the biomedical domain is a known issue [261, 262, 263]. Throughout this thesis, data standards have been used to alleviate some of the concerns - for example, there is no need for integration of multiple database schemas or languages. Another deliberate choice was to publish all codes and dataset on publicly available content management system, such as Sourceforge [264] and Google Code [265], and rely on the OCLC PURL infrastructure to rem-

edy disappearing URLs. It is also anticipated that consortium development in the context of the OBO Foundry will provide community support of resources, therefore decreasing their chance of vanishing. To help address some of those issues, as well as maintain the infrastructure described in Section 6.2, a new group has been established in June 2012, with mission to streamline the OBO Foundry operations and supports its coordinated maintenance. As part of this OBO Operations Committee (OBOFOC, `http://code.google.com/p/obo-foundry-operations-committee/`), a dedicated technical group aims at supporting the OBO global infrastructure as well as desiderata from the ontologies developers. As part of this group, I authored four documents describing the details of the systems currently deployed [266, 267, 268, 269]. More efforts need to be done to address legacy documentation, as well as consolidate existing infrastructure, for example by implementing backup/mirror systems in case of failure.

10.2.2 Evolution of the AERO

The AERO is open-access and available publicly at `http://purl.obolibrary.org/obo/aero`. Dr Jan Bonhoeffer from the Brighton Collaboration has expressed interest in translating the BCCDs within the ADVANCE (Accelerated development of vaccine benefit-risk collaboration in Europe) network which recently launched. A Brighton working group has also been created, and it is expected it will at a minimum keep the different interested parties in contact with respect to application of the AERO to remaining BCCDs. Following a meeting in Buffalo in June 2012, several ontology developers, including representatives from the OGMS, IDO, Ontology of Adverse Events (OAE), expressed interest in building a global infrastructure for all surveillance projects within the OBO Foundry. Other parties, such as members of the Network of Relevant Ontologies for Epidemiology consortium in charge of the Epidemiology Ontology [270], and representatives of the FDA Medical Device Safety division were also looking forward to a common representation of the medical interventions and following events. To that aim, the Medical Surveillance Ontology (MSrv, [271]) has been created. I had extensive discussions with the developers of the OAE, and while there is agreement with respect to integration of some very high-level terms, several issues haven't been addressed, and subsequently the MSrv is still in very early stages of development.

10.2.3 Implementation in reporting systems

In [15], completeness of the information recorded is identified as the limiting factor. The authors suggest that "One possible solution, which may allow any of the BCCDs to be applied, would be

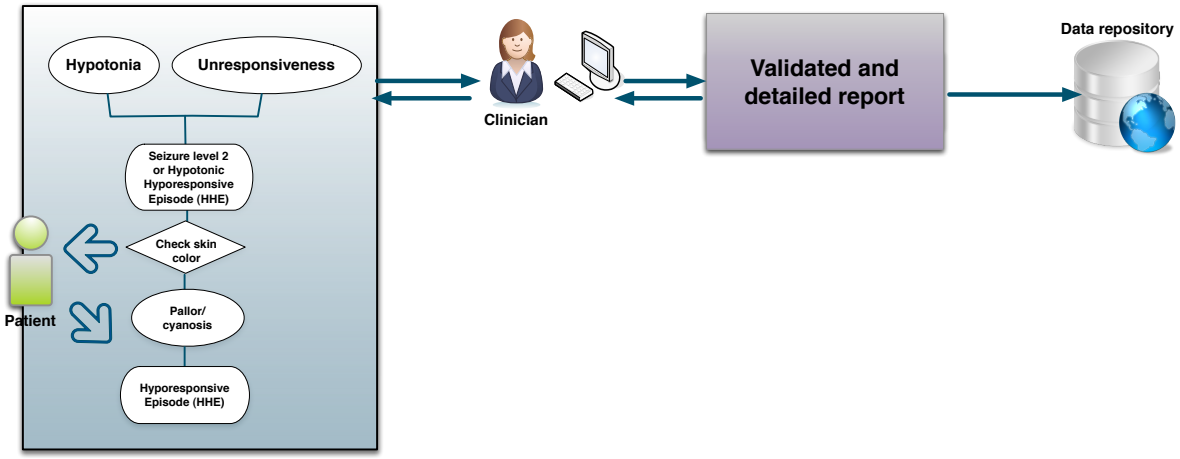


Figure 10.1: Diagnosis confirmation. An automated system can help confirm diagnosis at the time of data entry, by suggesting additional criteria to disambiguate diagnoses. In this case, observation of the patient skin color is enough information to determine if the event reported is a seizure or a hypotonic hypo-responsive episode.

to educate health care providers on what specific symptoms, signs and investigations should be captured." One application of the work done within this thesis is to enable the use of an ontology-based system at the time of data entry, which will increase data accuracy and completeness. For example, when the clinicians select "seizure" as adverse event, they will be offered a list of symptoms that may have manifested. By selecting the ones they did observe, the system will be able to confirm their diagnosis, potentially specifying it, such as assigning a level of certainty based on the Brighton case definition. The system will also be able to call the diagnosis into question, by warning that the set of events selected does not allow for unambiguous interpretation, such as shown on Figure 10.1. In the latter case, the system will also provide a list of such events that would allow determination. This will enable, at the time of data entry, clinicians to unambiguously refer to a specific set of symptoms, each carefully defined, and establish a diagnosis that remains linked to its associated symptoms. The adverse event will also be formally expressed, making it amenable to further querying for example for statistical analysis "what percentage of patients presented with motor manifestations?") at different levels of granularity (e.g., facilitating queries such as "what percentage of patients presented with tonic-clonic motor manifestations?")

This system not only addresses the concern that not all required signs and symptoms are being reported, but it could additionally check on the consistency of the reported information. For example, if the health care provider assesses an anaphylaxis diagnosis, the system can prompt them

to record required observations in the multiple systems required to be involved in such diagnosis, but also check that taken together they are consistent with the BCCD anaphylaxis. Enabling such interaction at the point of care would be beneficial for the reporting systems, as it may limit the number of back and forth required between the local organization and the national surveillance group, who often needs to requests more information (e.g., detailed medical records) to confirm diagnoses.

Additionally, two main barriers for adoption of the anaphylaxis BCCD were identified [15]: (1) not all signs and symptoms required for application of the BCCD are reported; and (2) those signs and symptoms are not consistently described and reported. While the use of a glossary promises to address point (2), the availability of a unique system that would solve both issues would be preferable. The ontology can both support reporting of required signs and symptoms and check their consistency, and can also offer help to the user via either textual definitions (which have currently been integrated in the AERO from the PHAC glossary) or even via their logical representation. Indeed, nothing precludes nesting of ontology terms. For example, fever can be a diagnosis when the fever BCCD is applied, but it can be a sign or symptom when the Seizure BCCD is applied.

10.2.4 Application to other guidelines and other domains

Formally expressing the signs and symptoms via the AERO allows for integration of multiple perspectives in health care. In Chapter 8, I showed that the AERO can be applied to the WHO Malaria guidelines. Additionally, considering that it is unlikely all systems will adopt the same guidelines worldwide, and that drugs are being shared across countries - aggregating adverse event information internationally is critical, as was first shown in the Thalidomide case [6], leading to the establishment of the WHO. With the AERO, users have the ability to encode their specific guideline, relying on common building blocks from the OGMS or the OMRE. A reasoner can then be used to classify automatically documents into one or several categories. For example, in the US, the CDC defines Influenza-like Illness (ILI) as “fever over 100 °F AND cough and/or sore throat”. In Canada, PHAC uses the definition “Acute onset of respiratory illness with fever and cough and with one or more of the following - sore throat, arthralgia, myalgia, or prostration which is likely due to influenza.” While both case definitions require fever and cough, the PHAC one goes further and requires an extra sign/symptom. Using the AERO, both guidelines can be encoded, and individual patients instances can be classified in one or more categories: all patients that are

CDC ILI are PHAC ILI (but the reverse is not true).

10.2.5 Data integration and text-mining

Very early work has been done in linking the VAERS dataset with other resources, thus fulfilling the last of the linked data principles, "Include links to other URIs in that useful information, so that more things can be discovered." In doing so, some issues arose, such as missing terms in the VO. Work is ongoing with VO developers to add the remaining information in their ontology. It is expected that this will enable more complex querying, such as "are there differences in the type of adverse events observed with different types of vaccines?". A limitation of my work is that only structured annotations (the MedDRA terms) were considered. However, each VAERS report also includes a textual part, which can be more or less detailed. Some work has been done to apply Natural Language Processing methods to analysis of adverse event reports [243]. Together with a private company, Seeker solutions (`http://www.seekersolutions.com`), a preliminary analysis of the textual content of VAERS reports was performed - results of which are attached in Appendix H. While theoretically promising, many hurdles stand in the way of properly exploiting the textual part of the reports. First, the content itself varies greatly in terms of length and quality. A number of medical abbreviations are used, and some reports are filled in foreign language (e.g., Spanish in the case of the VAERS). Also, it is unclear whether the text is comprehensive or not - are all the observed signs and symptoms reported? This is an issue I mentioned in Chapter 9, and that could be overcome with better reporting methods at the time of data entry.

Finally, it would be very interesting to pursue a combination of text-mining and data integration, which would leverage the content of the reports and the power of the Semantic Web. For example, if it was possible to extract names of the drugs that were used for treatment of patient, and which are often mentioned in the text, they could be linked with information from DrugBank [272]. Knowing a patient was treated with Benadryl, and via DrugBank that Benadryl is an anti-allergic agent, that could be inferred as supportive evidence (though weaker) for potential anaphylaxis.

10.3 Conclusion

This thesis forms a coherent body of work showing how existing biomedical knowledge can be encoded using formal representations. It details several resources I contributed to, and my involvement within the OBO Foundry to support interoperability of resources, and publication on

the Semantic Web. Using the pharmacovigilance domain, it demonstrates how ontologies can be used to improve standardization of knowledge, as well as automate some manual processes, such as classification of adverse event reports. Additionally, it proposes some ways ontologies could be practically implemented to improve the reporting process. Finally, this thesis achieves the goal of raising awareness in the clinical community: following my results, the Brighton Collaboration is moving towards providing an ontological representation of their existing and future guidelines. This will hopefully pave the way for other organizations to understand and rely on ontology-based applications.

Bibliography

[1] Government of Canada Panel on Reasearch Ethics. TCPS 22nd edition of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans - http://www.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/chapter2-chapitre2/#toc02-1a, Accessed Feb 2014. (Cited on page iv.)

[2] James A Singleton, Jenifer C Lloyd, Gina T Mootrey, Marcel E Salive, and Robert T Chen. An overview of the vaccine adverse event reporting system (VAERS) as a surveillance system. *Vaccine*, 17(22):2908–2917, 1999. (Cited on page 1.)

[3] A. Sinha, G. Hripcsak, and M. Markatou. Large datasets in biomedicine: a discussion of salient analytic issues. *Journal of the American Medical Informatics Association*, 16(6):759–767, 2009. (Cited on pages 1 and 13.)

[4] World Health Organization (WHO). The importance of pharmacovigilance: safety monitoring of medicinal products. *Geneva: World Health Organization*, pages 1–48, 2002. (Cited on page 7.)

[5] Bara Fintel, Athena T. Samaras, and Carias Edson. The thalidomide tragedy: lessons for drug safety and regulation - http://helix.northwestern.edu/article/thalidomide-tragedy-lessons-drug-safety-and-regulation, 2009. (Cited on page 7.)

[6] Frances O Kelsey. Thalidomide update: regulatory aspects. *Teratology*, 38(3):221–226, 1988. (Cited on pages 7 and 155.)

[7] Vaccine European New Integrated Collaboration Effort - http://venice.cineca.org/, June 2011. (Cited on page 7.)

[8] VENICE project. Final Report on the Survey on AEFI Monitoring Systems in Member States - http://venice.cineca.org/WP5_final_report.pdf, June 2011. (Cited on page 8.)

[9] US Food and Drug Administration. Adverse Event Reporting System Data - http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm, June 2011. (Cited on page 8.)

[10] US Food and Drug Administration. Vaccine Adverse Event Reporting System Data - http://vaers.hhs.gov/data/index, June 2011. (Cited on pages 8 and 9.)

[11] Gary H. Merrill. The MedDRA paradox. *AMIA Annual Symposium Proceedings*, 2008:470–474, 2008. (Cited on page 8.)

[12] Krischer J Richesson R, Fung K. Heterogeneous but standard coding systems for adverse events: Issues in achieving interoperability between apples and oranges. *Contemp Clin Trials*, 29, 2008. (Cited on page 8.)

[13] P. Mozzicato. Standardised MedDRA queries: their role in signal detection. *Drug Saf*, 30(7):617–619, 2007. (Cited on page 8.)

[14] June Almenoff, Joseph M Tonning, A Lawrence Gould, Ana Szarfman, Manfred Hauben, Rita Ouellet-Hellstrom, Robert Ball, Ken Hornbuckle, Louisa Walsh, Chuen Yee, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug safety*, 28(11):981–1007, 2005. (Cited on page 8.)

[15] Michael S. Gold, Jane Gidudu, Mich Erlewyn-Lajeunesse, and Barbara Law. Can the Brighton Collaboration case definitions be used to improve the quality of Adverse Event Following Immunization (AEFI) reporting?: Anaphylaxis as a case study. *Vaccine*, 28(28):4487 – 4498, 2010. (Cited on pages 8, 14, 16, 21, 123, 150, 153, and 155.)

[16] Canada Minister of Health. Clinical Safety Data Management Definitions and Standards for Expedited Reporting - `http://www.hc-sc.gc.ca/dhp-mps/prodpharma/applic-demande/guide-ld/ich/efficac/e2a-eng.php#a2A1`, June 2011. (Cited on page 8.)

[17] Public Health Agency of Canada. User Guide: Report of Adverse Events Following Immunization (AEFI) - `http://www.phac-aspc.gc.ca/im/pdf/AEFI-ug-gu-eng.pdf`, June 2011. (Cited on pages 9 and 10.)

[18] JS Goraya and VS Virdi. Bacille Calmette-Guérin lymphadenitis. *Postgraduate medical journal*, 78(920):327–329, 2002. (Cited on page 9.)

[19] Lis Halkieer-Lassen. Suppurative lymphadenitis following intradermal BCG vaccination of pre-school children. *Bull. Org. mond. Sante*, 12:143–167, 1955. (Cited on page 9.)

[20] Paul Hengster, J Schnapka, M Fille, and G Menardi. Occurrence of suppurative lymphadenitis after a change of BCG vaccine. *Archives of disease in childhood*, 67(7):952–955, 1992. (Cited on page 9.)

[21] Maurice Arthus. Injections repetees de serum de cheval chez le lapin. *Comptes Rendus des Seances de la Societe de Biologie et de ses Filiales*, pages 817–820, 1903. (Cited on page 9.)

[22] Danuta M Skowronski, Barbara Strauss, Gaston De Serres, Diane MacDonald, Stephen A Marion, Monika Naus, David M Patrick, and Perry Kendall. Oculo-respiratory syndrome: a new influenza vaccine-associated adverse event? *Clinical infectious diseases*, 36(6):705–713, 2003. (Cited on page 10.)

[23] Danuta M Skowronski, Gaston De Serres, Jacques Hebert, Donald Stark, Richard Warrington, Jane Macnabb, Ramak Shadmani, Louis Rochette, Diane MacDonald, David M Patrick, and Bernard Duval. Skin testing to evaluate oculo-respiratory syndrome (ORS) associated with influenza vaccination during the 2000–2001 season. *Vaccine*, 20(21):2713–2719, 2002. (Cited on page 10.)

[24] Philipp J Fritsche, Arthur Helbling, and Barbara K Ballmer-Weber. Vaccine hypersensitivity–update and overview. *Swiss Med Wkly*, 140(17-18):238–246, 2010. (Cited on page 10.)

[25] Leslie K. Ball, Geoffrey Evans, and Ann Bostrom. Risky business: Challenges in vaccine risk communication. *Pediatrics*, 101(3):453–458, 1998. (Cited on pages 11 and 13.)

[26] R.T. Chen, S.C. Rastogi, J.R. Mullen, S.W. Hayes, S.L. Cochi, J.A. Donlon, and S.G. Wassilak. The vaccine adverse event reporting system (VAERS). *Vaccine*, 12(6):542–550, 1994. (Cited on pages 11 and 137.)

[27] Gregory A. Poland and Robert M. Jacobson. The age-old struggle against the antivaccinationists. *New England Journal of Medicine*, 364(2):97–99, 2011. (Cited on page 12.)

[28] Aj Wakefield, Sh Murch, A Anthony, J Linnell, Dm Casson, M Malik, M Berelowitz, Ap Dhillon, Ma Thomson, P Harvey, A Valentine, Se Davies, and Ja Walker-Smith. RETRACTED: ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351:637–641, February 1998. (Cited on page 12.)

[29] Jason M. Glanz, David L. McClure, David J. Magid, Matthew F. Daley, Eric K. France, Daniel A. Salmon, and Simon J. Hambidge. Parental refusal of pertussis vaccination is associated with an increased risk of pertussis infection in children. *Pediatrics*, 123(6):1446 –1451, June 2009. (Cited on page 13.)

[30] Jerry Avorn and Daniel H. Solomon. Cultural and economic factors that (mis)shape antibiotic use: The nonpharmacologic basis of therapeutics. *Annals of Internal Medicine*, 133(2):128–135, 2000. (Cited on page 13.)

[31] Fraser Health. Measles cluster in Fraser East - `http://www.fraserhealth.ca/about_us/media_centre/news_releases/2013-news-releases/measles-cluster-in-fraser-east`, September Accessed Sep 2013. (Cited on page 13.)

[32] CNN. U.S. measles cases in 2013 may be most in 17 years - `http://www.cnn.com/2013/09/12/health/worst-measles-year/index.html`, September Accessed Sep 2013. (Cited on page 13.)

[33] F. Varricchio, J. Iskander, F. Destefano, R. Ball, R. Pless, M.M. Braun, and R.T. Chen. Understanding vaccine safety information from the vaccine adverse event reporting system. *The Pediatric infectious disease journal*, 23(4):287–294, 2004. (Cited on page 13.)

[34] US Center for Disease Control and Food Drug Administration. VAERS Reporting form - `https://vaers.hhs.gov/resources/vaers_form.pdf`, Accessed Feb 2014. (Cited on page 13.)

[35] Katrin S Kohl, Jan Bonhoeffer, M Miles Braun, Robert T Chen, Philippe Duclos, Harald Heijbel, Ulrich Heininger, and Elisabeth Loupi. The Brighton Collaboration: Creating a global standard for case definitions (and guidelines) for adverse events following immunization. *AHRQ: Advances in Patient Safety. Concepts and Methodology. Rockville, AHRQ*, 2:87–102, 2005. (Cited on page 14.)

[36] The Brighton Collaboration. `http://www.brightoncollaboration.org`, June Accessed Jun 2011. (Cited on pages 14 and 15.)

[37] Australasian Society of Clinical Immunology and Allergy - `http://www.allergy.org.au`, December 2012. (Cited on page 15.)

[38] National Institute for Health and Clinical Excellence - `http://www.nice.org.uk`, December 2012. (Cited on page 15.)

[39] World Allergy Organization - `http://www.worldallergy.org`, December 2012. (Cited on page 15.)

[40] J. Bonhoeffer, K. Kohl, R. Chen, P. Duclos, H. Heijbel, U. Heininger, T. Jefferson, and E. Loupi. The Brighton Collaboration: addressing the need for standardized case definitions of adverse events following immunization (AEFI). *Vaccine*, 21(3):298–302, 2002. (Cited on page 14.)

[41] The Brighton Collaboration. ABC tool - available at `http://legacy.brightoncollaboration.org/intranet/en/index/aefi_classifyer.html`. registration required., June 2010. (Cited on page 16.)

[42] Public Health Agency of Canada `http://www.phac-aspc.gc.ca/im/aefi-essi_guide/page3-eng.php#sec9`. User Guide: Report of Adverse Events Following Immunization (AEFI), Accessed Dec 2013. (Cited on pages 16 and 21.)

[43] Rachel Regier, Rupali Gurjar, and Roberto A Rocha. A clinical rule editor in an electronic medical record setting: development, design, and implementation. In *AMIA Annual Symposium Proceedings*, volume 2009, page 537. American Medical Informatics Association, 2009. (Cited on page 16.)

[44] Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in bioinformatics*, 6(3):239–251, 2005. (Cited on pages 16 and 22.)

[45] Jens U. Ruggeberg, Michael S. Gold, Jose-Maria Bayas, Michael D. Blum, Jan Bonhoeffer, Sheila Friedlander, Glacus de Souza Brito, Ulrich Heininger, Babatunde Imoukhuede, Ali Khamesipour, Michel Erlewyn-Lajeunesse, Susana Martin, Mika Makela, Patricia Nell, Vitali Pool, and Nick Simpson. Anaphylaxis: Case definition and guidelines for data collection, analysis, and presentation of immunization safety data. *Vaccine*, 25(31):5675 – 5684, 2007. (Cited on page 16.)

[46] Public Health Agency of Canada `http://www.phac-aspc.gc.ca/publicat/cig-gci/p02-01-eng.php`. Canadian immunization guide, Accessed Dec 2013. (Cited on page 17.)

[47] J. P. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. Repeatability of published microarray gene expression analyses. *Nature genetics*, 41(2):149–155, Feb 2009. (Cited on page 21.)

[48] C. J. Penkett and J. Bahler. Navigating public microarray databases. *Comparative and Functional Genomics*, 5(6-7):471–479, 2004. (Cited on page 21.)

[49] R. G. Cote, P. Jones, L. Martens, R. Apweiler, and H. Hermjakob. The Ontology Lookup Service: more data and better tools for controlled vocabulary queries. *Nucleic acids research*, 36(Web Server issue):W372–6, Jul 1 2008. (Cited on pages 21 and 105.)

[50] James J Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, 37(4-5):394, 1998. (Cited on pages 21 and 23.)

[51] James J Cimino. In defense of the Desiderata. *Journal of biomedical informatics*, 39(3):299–306, 2006. (Cited on page 22.)

[52] D. L. Rubin, N. H. Shah, and N. F. Noy. Biomedical ontologies: a functional perspective. *Briefings in bioinformatics*, 9(1):75–90, Jan 2008. (Cited on page 22.)

[53] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A. harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. (Cited on pages 22, 24, 54, and 97.)

[54] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005. (Cited on pages 22, 49, and 57.)

[55] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007. (Cited on pages 22, 24, 25, 49, 76, and 124.)

[56] PubMed Home - `http://www.ncbi.nlm.nih.gov/pubmed/`. (Cited on pages 22, 48, and 56.)

[57] Alan Ruttenberg, Jonathan A Rees, Matthias Samwald, and M Scott Marshall. Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in bioinformatics*, 10(2):193–204, 2009. (Cited on pages 22, 23, 28, 80, and 90.)

[58] Jyotishman Pathak, Thomas M Johnson, and Christopher G Chute. Survey of modular ontology techniques and their applications in the biomedical domain. *Integrated Computer-Aided Engineering*, 16(3):225–242, 2009. (Cited on page 22.)

[59] Bernardo Cuenca Grau, Ian Horrocks, Yevgeny Kazakov, and Ulrike Sattler. Modular reuse of ontologies: Theory and practice. *J. Artif. Intell. Res.(JAIR)*, 31:273–318, 2008. (Cited on page 22.)

[60] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51, 2007. (Cited on pages 23, 28, 60, 75, and 147.)

[61] Olivier Bodenreider et al. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*, 47:67–79, 2008. (Cited on page 23.)

[62] Judith A Blake and Carol J Bult. Beyond the data deluge: data integration and bio-ontologies. *Journal of biomedical informatics*, 39(3):314–320, 2006. (Cited on page 23.)

[63] Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7, 2004. (Cited on pages 23 and 97.)

[64] Nicolas Le Novere, Benjamin Bornstein, Alexander Broicher, Mélanie Courtot, Marco Donizelli, Harish Dharuri, Lu Li, Herbert Sauro, Maria Schilstra, Bruce Shapiro, et al. Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic acids research*, 34(suppl 1):D689–D691, 2006. (Cited on page 23.)

[65] Robert Stevens, Patricia Baker, Sean Bechhofer, Gary Ng, Alex Jacoby, Norman W Paton, Carole A Goble, and Andy Brass. Tambis: transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2):184–186, 2000. (Cited on page 23.)

[66] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–716, 2008. (Cited on pages 23, 28, 105, and 107.)

[67] Neurocommons sparql endpoint - `http://sparql.neurocommons.org/`. (Cited on pages 23 and 86.)

[68] LOD SPARQL Endpoint - `http://lod.openlinksw.com/sparql`. (Cited on pages 23 and 28.)

[69] Gwenaëlle Marquet, Olivier Dameron, Stephan Saikali, Jean Mosser, and Anita Burgun. Grading gene tumors using OWL-DL and NCI Thesaurus. In *AMIA Annual Symposium Proceedings*, volume 2007, page 508. American Medical Informatics Association, 2007. (Cited on page 23.)

[70] Barry Smith. Ontology (science). In *FOIS*, pages 21–35, 2008. (Cited on page 23.)

[71] Midori A Harris, Jennifer I Deegan, Jane Lomax, Michael Ashburner, Susan Tweedie, Seth Carbon, Suzanna Lewis, Chris Mungall, John Day-Richter, Karen Eilbeck, et al. The gene ontology project in 2008. *Nucleic Acids Res*, 36:D440–D444, 2008. (Cited on page 24.)

[72] P De Matos, M Ennis, M Darsow, M Guedj, K Degtyarenko, and R Apweiler. ChEBI-chemical entities of biological interest. *Nucleic Acids Research*, 2006. (Cited on page 24.)

[73] Patricia L Whetzel, Ryan R Brinkman, Helen C Causton, Liju Fan, Dawn Field, Jennifer Fostel, Gilberto Fragoso, Tanya Gray, Mervi Heiskanen, Tina Hernandez-Boussard, et al. Development of FuGO: an ontology for functional genomics investigations. *OMICS: A journal of integrative biology*, 10(2):199–204, 2006. (Cited on page 24.)

[74] Patricia L Whetzel, Helen Parkinson, Helen C Causton, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Laurence Game, Mervi Heiskanen, Norman Morrison, Philippe Rocca-Serra, et al. The mged ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22(7):866–873, 2006. (Cited on page 24.)

[75] Larisa N Soldatova and Ross D King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006. (Cited on page 24.)

[76] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, et al. The automation of science. *Science*, 324(5923):85–89, 2009. (Cited on page 24.)

[77] Susanna-Assunta Sansone, Daniel Schober, Helen J Atherton, Oliver Fiehn, Helen Jenkins, Philippe Rocca-Serra, Denis V Rubtsov, Irena Spasic, Larisa Soldatova, Chris Taylor, et al. Metabolomics standards initiative: ontology working group work in progress. *Metabolomics*, 3(3):249–256, 2007. (Cited on page 24.)

[78] B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46, 2005. (Cited on pages 25, 32, and 124.)

[79] P. Grenon, B. Smith, and L. Goldberg. Biodynamic ontology: applying BFO in the biomedical domain. *Studies in health technology and informatics*, 102:20–38, 2004. (Cited on pages 25, 34, 49, and 76.)

[80] Alexander C Yu. Methods in biomedical ontology. *Journal of biomedical informatics*, 39(3):252–266, 2006. (Cited on page 25.)

[81] Alan L Rector et al. Clinical terminology: why is it so hard? *Methods of information in medicine*, 38(4/5):239–252, 1999. (Cited on page 25.)

[82] Yongqun He, Lindsay Cowell, Alexander D Diehl, HL Mobley, Bjoern Peters, Alan Ruttenberg, Richard H Scheuermann, Ryan R Brinkman, Mélanie Courtot, Chris Mungall, Zuoshuang Xiang, Fang Chen Chen, Thomas Todd, Lesley Colby, Howard Rush Rush, Trish Whetzel, Mark A. Musen, Brian D. Athey, Gilbert S. Omenn Omenn, and Barry Smith. VO: vaccine ontology. In *The 1st International Conference on Biomedical Ontology (ICBO 2009) Nature Precedings*, pages 24–26, 2009. (Cited on page 25.)

[83] OBI consortium. Ontology for Biomedical Investigations (OBI) - `http://purl.obolibrary.org/obo/obi`, June 2011. (Cited on pages 25, 75, and 124.)

[84] Mélanie Courtot, Chris Mungall, Ryan R. Brinkman, and Alan Ruttenberg. Building the OBO Foundry - one policy at a time. In *Proceedings of the International Conference on Biomedical Ontology (ICBO2011)*, 2011. (Cited on page 25.)

[85] The Basic Formal Ontology (BFO) - `http://www.ifomis.org/bfo/`, Accessed Dec 2013. (Cited on pages 25 and 47.)

[86] Chris F Taylor, Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, Michael Ashburner, Catherine A Ball, Pierre-Alain Binz, Molly Bogue, Tim Booth, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature biotechnology*, 26(8):889–896, 2008. (Cited on page 25.)

[87] Amit P Sheth and James A Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys (CSUR)*, 22(3):183–236, 1990. (Cited on page 26.)

[88] Gomer Thomas, Glenn R Thompson, Chin-Wan Chung, Edward Barkmeyer, Fred Carter, Marjorie Templeton, Stephen Fox, and Berl Hartman. Heterogeneous distributed database systems for production use. *ACM Computing Surveys (CSUR)*, 22(3):237–266, 1990. (Cited on page 26.)

[89] Richard Hull. Managing semantic heterogeneity in databases: a theoretical prospective. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 51–61. ACM, 1997. (Cited on page 26.)

[90] Alon Y Halevy. Answering queries using views: A survey. *The VLDB Journal*, 10(4):270–294, 2001. (Cited on page 26.)

[91] Dan Brickley and Ramanathan V Guha. Resource Description Framework (RDF) Schema Specification 1.0: W3C Candidate Recommendation 27 March 2000 - `http://www.w3.org/TR/2000/CR-rdf-schema-20000327/`, Accessed Nov 2013. (Cited on pages 26 and 105.)

[92] World Wide Web Consortium (W3C). OWL Web Ontology Language Guide, 02/10/ 2004. (Cited on pages 26 and 60.)

[93] Leora Morgenstern, Chris Welty, and Harold Boley. RIF Primer. *World-Wide Web Consortium*, 2010. (Cited on page 26.)

[94] SPARQL Query Language for RDF - `http://www.w3.org/TR/rdf-sparql-query/`. (Cited on pages 26, 29, 78, and 86.)

[95] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001. (Cited on page 27.)

[96] Tim Berners-Lee `http://www.w3.org/DesignIssues/LinkedData.html`. Linked Data, Accessed Dec 2013. (Cited on page 27.)

[97] UniProt Consortium. The universal protein resource (UniProt). *Nucleic acids research*, 36(Database issue):D190–5, Jan 2008. (Cited on pages 28 and 104.)

[98] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. (Cited on page 28.)

[99] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, Jan 1 2000. (Cited on page 28.)

[100] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, J. U. Pontius, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 33(Database issue):D39–45, Jan 1 2005. (Cited on page 28.)

[101] Rob Shearer, Boris Motik, and Ian Horrocks. HermiT: A Highly-Efficient OWL Reasoner. In *OWLED*, volume 432, 2008. (Cited on pages 28, 139, and 147.)

[102] Dmitry Tsarkov and Ian Horrocks. Fact++ description logic reasoner: System description. In *Automated reasoning*, pages 292–297. Springer, 2006. (Cited on pages 28 and 60.)

[103] OpenLink Virtuoso: Open-Source Edition, `http://virtuoso.openlinksw.com/wiki/main/Main/`. (Cited on pages 28, 90, and 147.)

[104] Clark & Parsia. Stardog - the RDF database - `http://stardog.com`, Accessed Dec 2013. (Cited on page 28.)

[105] Atanas Kiryakov, Damyan Ognyanov, and Dimitar Manov. OWLIM–a pragmatic semantic repository for OWL. In *Web Information Systems Engineering–WISE 2005 Workshops*, pages 182–192. Springer, 2005. (Cited on page 28.)

[106] Aduna. The Sesame triplestore - `http://www.openrdf.org/index.jsp`, August Accessed Aug 2013. (Cited on pages 29 and 139.)

[107] Jeen Broekstra, Michel Klein, Stefan Decker, Dieter Fensel, Frank Van Harmelen, and Ian Horrocks. Enabling knowledge representation on the web by extending rdf schema. *Computer networks*, 39(5):609–634, 2002. (Cited on page 30.)

[108] Ronald J Brachman and Hector J Levesque. The tractability of subsumption in frame-based description languages. In *AAAI*, volume 84, pages 34–37, 1984. (Cited on page 30.)

[109] Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The Even More Irresistible SROIQ. *KR*, 6:57–67, 2006. (Cited on pages 30 and 76.)

[110] M. Horridge, N. Drummond, J. Goodwin, A. Rector, R. Stevens, and H.H. Wang. The Manchester OWL Syntax. In Bernardo Cuenca Grau, Pascal Hitzler, Conor Shankey, and Evan Wallace, editors, *Proceedings of OWL Experiences and Directions Workshop (OWLED2006)*, 2006. (Cited on pages 30, 50, 66, 125, 129, and 141.)

[111] Mike Bergman. Thinking 'Inside the Box' with Description Logics - `http://www.mkbergman.com/466/thinking-inside-the-box-with-description-logics/`, Accessed Dec 2013. (Cited on page 30.)

[112] Franz Baader. *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2003. (Cited on page 30.)

[113] Boris Motik, Ian Horrocks, and Ulrike Sattler. Bridging the gap between OWL and relational databases. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(2):74–89, 2009. (Cited on page 31.)

[114] Sean Bechhofer, Frank Van Harmelen, Jim Hendler, Ian Horrocks, Deborah L McGuinness, Peter F Patel-Schneider, Lynn Andrea Stein, et al. OWL web ontology language reference. *W3C recommendation*, 10:2006–01, 2004. (Cited on pages 31, 32, and 49.)

[115] Samantha Bail. Common reasons for ontology inconsistency - `http://ontogenesis.knowledgeblog.org/1343`, Accessed Dec 2013. (Cited on page 32.)

[116] Johan Lauwereyns, Katsumi Watanabe, Brian Coe, and Okihide Hikosaka. A neural correlate of response bias in monkey caudate nucleus. *Nature*, 418(6896):413–417, 2002. (Cited on pages 34 and 35.)

[117] Christoph Hock, Uwe Konietzko, Andreas Papassotiropoulos, Axel Wollmer, Johannes Streffer, Ruth C von Rotz, Gabriela Davey, Eva Moritz, and Roger M Nitsch. Generation of antibodies specific for β-amyloid by vaccination of patients with alzheimer disease. *Nature medicine*, 8(11):1270–1275, 2002. (Cited on page 35.)

[118] William J Bug, Giorgio A Ascoli, Jeffrey S Grethe, Amarnath Gupta, Christine Fennema-Notestine, Angela R Laird, Stephen D Larson, Daniel Rubin, Gordon M Shepherd, Jessica A Turner, et al. The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, 6(3):175–194, 2008. (Cited on page 35.)

[119] The Vaccine Ontology - `http://www.violinet.org/vaccineontology/`, June 2011. (Cited on pages 37, 82, and 124.)

[120] Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37(suppl 1):D5–D15, 2009. (Cited on page 38.)

[121] Zuoshuang Xiang, Thomas Todd, Kim P Ku, Bethany L Kovacic, Charles B Larson, Fang Chen, Andrew P Hodges, Yuying Tian, Elizabeth A Olenzek, Boyang Zhao, et al. VIOLIN: vaccine investigation and online information network. *Nucleic acids research*, 36(suppl 1):D923–D928, 2008. (Cited on pages 41, 51, and 55.)

[122] Ryan R Brinkman, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, et al. Modeling biomedical experimental processes with OBI. *J Biomed Semantics*, 1(Suppl 1):S7, 2010. (Cited on pages 44, 50, and 55.)

[123] Gerhardt G Schurig, R Martin Roop, T Bagchi, S Boyle, D Buhrman, and N Sriranganathan. Biological

properties of RB51; a stable rough strain of *Brucella abortus*. *Veterinary microbiology*, 28(2):171–188, 1991. (Cited on pages 44 and 46.)

[124] M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg. MIREOT: The minimum information to reference an external ontology term. *Applied Ontology*, 6(1):23–33, 2011. (Cited on pages 47, 55, and 63.)

[125] The Information Artifact Ontology (IAO) - `http://purl.obolibrary.org/obo/iao`, June 2011. (Cited on pages 47, 49, 79, 92, and 124.)

[126] The Protégé Ontology Editor and Knowledge Acquisition System, `http://protege.stanford.edu/`. (Cited on pages 49, 60, 74, 75, 97, 114, and 147.)

[127] M. Musen, N. Shah, N. Noy, B. Dai, M. Dorf, N. Griffith, J. D. Buntrock, C. Jonquet, M. J. Montegut, and D. L. Rubin. Bioportal: Ontologies and data resources with the click of a mouse. *AMIA ...Annual Symposium proceedings / AMIA Symposium.AMIA Symposium*, pages 1223–1224, 2008. (Cited on page 49.)

[128] G. V. Gkoutos, E. C. Green, A. M. Mallon, J. M. Hancock, and D. Davidson. Using ontologies to describe mouse phenotypes. *Genome Biol*, 6(1):R8, 2005. 1465-6914 Journal Article. (Cited on pages 49, 50, 74, 75, and 97.)

[129] Cornelius Rosse and José LV Mejino Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Journal of biomedical informatics*, 36(6):478–500, 2003. (Cited on pages 49 and 69.)

[130] National Institutes of Health National Center for Biotechnology Information (NCBI), National Library of Medicine. The NCBI Entrez Taxonomy Homepage. (Cited on pages 49 and 55.)

[131] IDO consortium. The Infectious Disease Ontology - `http://www.infectiousdiseaseontology.org/`, December Accessed Dec 2009. (Cited on pages 50 and 82.)

[132] The OGMS developers group. Ontology for General Medical Science (OGMS) - `http://code.google.com/p/ogms/`, Accessed Jun 2011. (Cited on pages 51 and 124.)

[133] Alan L Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including owl. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 121–128. ACM, 2003. (Cited on page 52.)

[134] Ali M Harandi, Gwyn Davies, and Ole F Olesen. Vaccine adjuvants: scientific challenges and strategic initiatives. *Expert Review of Vaccines*, 2009. (Cited on page 53.)

[135] Seth Carbon, Amelia Ireland, Christopher J Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, et al. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009. (Cited on pages 54, 105, and 107.)

[136] FLU consortium. The Influenza Ontology - `https://sourceforge.net/projects/influenzo/`, December Accessed Dec 2009. (Cited on pages 55 and 82.)

[137] Henry J Lowe and G Octo Barnett. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA: the journal of the American Medical Association*, 271(14):1103–1108, 1994. (Cited on page 56.)

[138] Zuoshuang Xiang, Wenjie Zheng, and Yongqun He. BBP: Brucella genome annotation with literature mining and curation. *BMC bioinformatics*, 7(1):347, 2006. (Cited on page 56.)

[139] Olivier Bodenreider and Robert Stevens. Bio-ontologies: current trends and future directions. *Briefings in bioinformatics*, 7(3):256–274, 2006. (Cited on page 60.)

[140] Mikel Egaña Aranguren, Erick Antezana, Martin Kuiper, and Robert Stevens. Ontology design patterns for bio-ontologies: a case study on the cell cycle ontology. *BMC bioinformatics*, 9(Suppl 5):S1, 2008. (Cited on page 60.)

[141] Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette Ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache. *The Semantic Web: Research and Applications: 7th European Semantic Web Conference, ESW 2010, Heraklion, Crete, Greece, May 30-June 3, 2010, Proceedings*, volume 2. Springer, 2010. (Cited on page 60.)

[142] Oscar Corcho, Catherine Roussey, LM Vilches-Blázquez, and Iván Perez Dominguez. Pattern-based OWL ontology debugging guidelines. In *OWLED*, 2009. (Cited on page 60.)

[143] Luigi Iannone, Mikel Egaña Aranguren, Alan L Rector, and Robert Stevens. Augmenting the expressivity of the ontology pre-processor language. In *OWLED*, volume 432, 2008. (Cited on pages 60 and 65.)

[144] The Bio Investigation Index. BII: The Bio Investigation Index. - `http://www.ebi.ac.uk/bioinvindex/home.seam`, Accessed Nov 2013. (Cited on page 61.)

[145] Dawn Field, Susanna-Assunta Sansone, Amanda Collis, Tim Booth, Peter Dukes, Susan K. Gregurick, Karen Kennedy, Patrik Kolar, Eugene Kolker, Mary Maxon, Sian Millard, Alexis-Michel Mugabushaka, Nicola Perrin, Jacques E. Remacle, Karin Remington, Philippe Rocca-Serra, Chris F. Taylor, Mark Thorley, Bela Tiwari, and John Wilbanks. 'Omics Data Sharing. *Science*, 326(5950):234–236, 2009. (Cited on pages 61 and 68.)

[146] Bjoern Peters and Alessandro Sette. Integrating epitope data into the emerging web of biomedical knowledge resources. *Nature Reviews Immunology*, 7(6):485–490, 2007. (Cited on page 61.)

[147] Burke Squires, Catherine Macken, Adolfo Garcia-Sastre, Shubhada Godbole, Jyothi Noronha, Victoria Hunt, Roger Chang, Christopher N Larsen, Ed Klem, Kevin Biersack, et al. BioHealthBase: informatics support in the elucidation of influenza virus host–pathogen interactions and virulence. *Nucleic acids research*, 36(suppl 1):D497–D503, 2008. (Cited on page 61.)

[148] Jay Kola. ExcelImport - co-ode-owl-plugins - Get data from a spreadsheet into your ontology - `http://code.google.com/p/co-ode-owl-plugins/wiki/ExcelImport`, Accessed Nov 2013. (Cited on page 65.)

[149] Martin J O'Connor, Christian Halaschek-Wiener, and Mark A Musen. M2: A Language for Mapping Spreadsheets to OWL. In *OWLED*, 2010. (Cited on page 66.)

[150] Matthew Horridge and Sean Bechhofer. The OWLAPI: A Java API for OWL ontologies. *Semantic Web*, 2(1):11–21, 2011. (Cited on pages 66, 90, 113, 139, and 147.)

[151] Philippe Rocca-Serra. Quick Term Templates - OBI Ontology - `http://obi-ontology.org/page/Quick_Term_Templates#Processing_the_Analyte_Assay_Template_using_Mapping_Master:_the_procedure_from_start_to_finish`, Accessed Nov 2013. (Cited on page 67.)

[152] International Union of Pure and Applied Chemistry. Subcommittee on Nomenclature, Properties, and Units in Laboratory Medicine - `http://old.iupac.org/divisions/VII/VII.C.1/index.html`, Accessed Nov 2013. (Cited on page 67.)

[153] Eagle i consortium. eagle-i - `https://www.eagle-i.net`, Accessed Nov 2013. (Cited on pages 68 and 118.)

[154] E. Maguire, P. Rocca-Serra, and S. Sansone. ISA Infrastructure - isacreator. - `http://isatab.sourceforge.net`, Accessed Nov 2013. (Cited on page 68.)

[155] Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl 1):D344–D350, 2008. (Cited on pages 69 and 76.)

[156] Georgios V Gkoutos, Paul N Schofield, and Robert Hoehndorf. The Units Ontology: a tool for integrating units of measurement in science. *Database: The Journal of Biological Databases and Curation*, 2012, 2012. (Cited on pages 69 and 81.)

[157] Darren A Natale, Cecilia N Arighi, Winona C Barker, Judith Blake, Ti-Cheng Chang, Zhangzhi Hu, Hongfang Liu, Barry Smith, and Cathy H Wu. Framework for a protein ontology. *BMC bioinformatics*, 8(Suppl 9):S1, 2007. (Cited on page 69.)

[158] Zuoshuang Xiang, Yu Lin, and Yongqun He. Ontorat web server for automatic generation and annotations of new ontology terms. In *International conference on biomedical Ontology (ICBO)*, 2012. (Cited on page 69.)

[159] John Day-Richter. The OBO Flat File Format Specification, version 1.2 - `http://www.geneontology.org/GO.format.obo-1_2.shtml`, Accessed Nov 2013. (Cited on page 71.)

[160] Web Ontology Language (OWL), `http://www.w3.org/2004/OWL/`. (Cited on pages 71, 75, 139, and 147.)

[161] OBI Ontology, `http://purl.obolibrary.org/obo/obi`, June 2011. (Cited on pages 71 and 125.)

[162] W3C. Simple Knowledge Organization System (SKOS) - `http://www.w3.org/TR/2009/REC-skos-reference-20090818/`, Accessed Nov 2013. (Cited on page 72.)

[163] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set - `http://dublincore.org/documents/dces/`, Accessed Nov 2013. (Cited on page 72.)

[164] Barry Smith. Beyond concepts: ontology as reality representation. In *Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)*, pages 73–84, 2004. (Cited on page 72.)

[165] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, 32(90001):D258–D261, 01/01/ 2004. (Cited on pages 73 and 75.)

[166] GO consortium. GO editorial style guide - `http://www.geneontology.org/GO.contents.doc.shtml`, October Accessed Oct 2013. (Cited on pages 73 and 76.)

[167] Martin Hepp. Goodrelations: An ontology for describing products and services offers on the web. In Aldo Gangemi and Jerome Euzenat, editors, *Knowledge Engineering: Practice and Patterns*, volume 5268 of *Lecture Notes in Computer Science*, pages 329–346. Springer Berlin / Heidelberg, 2008. (Cited on page 74.)

[168] UniProt Consortium et al. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research*, 41(D1):D43–D47, 2013. (Cited on page 74.)

[169] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources

of the national center for biotechnology information. *Nucleic acids research*, 39(suppl 1):D38–D51, 2011. (Cited on page 75.)

[170] C. Golbreich, S. Zhang, and O. Bodenreider. The foundational model of anatomy in OWL: Experience and perspectives. *Web semantics (Online)*, 4(3):181–195, 2006. (Cited on page 75.)

[171] Bernardo Cuenca Grau, Ian Horrocks, Yevgeny Kazakov, and Ulrike Sattler. Ontology reuse: Better safe than sorry. *Description Logics*, 250, 2007. (Cited on page 75.)

[172] Bernardo Cuenca Grau, Ian Horrocks, Yevgeny Kazakov, and Ulrike Sattler. Extracting modules from ontologies: A logic-based approach. In Heiner Stuckenschmidt and Stefano Spaccapietra, editors, *Ontology Modularization*. Springer, 2008. (Cited on pages 75 and 85.)

[173] Melissa A Haendel, Fabian Neuhaus, David Osumi-Sutherland, Paula M Mabee, Jos LV Mejino Jr, Chris J Mungall, and Barry Smith. CARO–the common anatomy reference ontology. In *Anatomy Ontologies for Bioinformatics*, pages 327–349. Springer, 2008. (Cited on page 75.)

[174] Bernardo Cuenca Grau, Ian Horrocks, Yevgeny Kazakov, and Ulrike Sattler. Just the right amount: extracting modules from ontologies. In *Proceedings of the 16th international conference on World Wide Web*, pages 717–726. ACM, 2007. (Cited on pages 75 and 122.)

[175] Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, Ulrike Sattler, Thomas Schneider, and Rafael Berlanga. Safe and economic re-use of ontologies: A logic-based methodology and tool support. In *The Semantic Web: Research and Applications*, pages 185–199. Springer, 2008. (Cited on pages 75 and 122.)

[176] Julian Seidenberg and Alan Rector. Web ontology segmentation: analysis, classification and use. In *Proceedings of the 15th international conference on World Wide Web*, pages 13–22. ACM, 2006. (Cited on pages 75, 76, and 85.)

[177] OBI scripts - `http://obi.svn.sourceforge.net/viewvc/obi/trunk/src/tools/`. (Cited on page 77.)

[178] OBI consortium. SPARQL queries template file - `http://obi.svn.sourceforge.net/svnroot/obi/trunk/src/tools/build/external-templates.txt`, December Accessed Dec 2009. (Cited on page 79.)

[179] Science Commons. Neurocommons OBO SPARQL endpoint - `http://sparql.obo.neurocommons.org/`, December Accessed Dec 2009. (Cited on page 80.)

[180] Jonathan Bard, Seung Y Rhee, and Michael Ashburner. An ontology for cell types. *Genome biology*, 6(2):R21, 2005. (Cited on page 80.)

[181] Dave Beckett and Brian McBride. RDF/XML syntax specification (revised) - `www.w3.org/TR/REC-rdf-syntax/?`, 2004. (Cited on pages 83, 86, and 105.)

[182] Zuoshuang Xiang, Mélanie Courtot, Ryan R Brinkman, Alan Ruttenberg, and Yongqun He. Ontofox: web-based support for ontology reuse. *BMC research notes*, 3(1):175, 2010. (Cited on page 84.)

[183] Roman Kontchakov, Luca Pulina, Ulrike Sattler, Thomas Schneider, Petra Selmer, Frank Wolter, and Michael Zakharyaschev. Minimal Module Extraction from DL-Lite Ontologies Using QBF Solvers. In *IJCAI*, volume 9, pages 836–841, 2009. (Cited on page 85.)

[184] Natalya F Noy and Mark A Musen. Specifying ontology views by traversal. In *The Semantic Web–ISWC 2004*, pages 713–725. Springer, 2004. (Cited on page 85.)

[185] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009. (Cited on page 103.)

[186] W3C OWL Working Group and others. OWL 2 Web Ontology Language document overview - `http://www.w3.org/TR/2009/REC-owl2-overview-20091027/`, 2009. (Cited on page 105.)

[187] Prud'hommeaux E and Seaborne A. Resource Description Framework (RDF) / W3C Semantic Web Activity - `http://www.w3.org/RDF/`. (Cited on page 105.)

[188] James Clark. XSL transformations (XSLT) - `http://www.w3.org/TR/xslt`, Accessed Nov 2013. (Cited on pages 105 and 109.)

[189] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl 2):W541–W545, 2011. (Cited on pages 105 and 106.)

[190] CO-ODE project. Ontology-browser: An OWL Ontology and RDF (Linked Open Data) Browser - `http://code.google.com/p/ontology-browser/`, Accessed Nov 2013. (Cited on pages 105 and 106.)

[191] Ontotext. Linked Life Data - `http://linkedlifedata.com/resource/geneontology/id/GO:0008150`, Accessed Nov 2013. (Cited on pages 105 and 107.)

[192] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia-A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009. (Cited on pages 105 and 108.)

[193] Steven Vercruysse, Aravind Venkatesan, and Martin Kuiper. OLSVis: an animated, interactive visual browser for bio-ontologies. *BMC bioinformatics*, 13(1):116, 2012. (Cited on page 106.)

[194] Compare http://bioportal.bioontology.org/ontologies/47893/?p=terms&conceptid=obo%3AOBI_0001705 to http://purl.obolibrary.org/obo/OBI_0001705, Accessed Dec 2012. (Cited on page 106.)

[195] BioPortal RDF retrieval of a SKOS term: http://rest.bioontology.org/bioportal/rdf/47893/?conceptid=http://purl.obolibrary.org/obo/OBI_0001705&apikey=6551953c-31bb-4189-91b5-f0733a251c61, Accessed Jan 2013. (Cited on page 106.)

[196] NCBO BioPortal. BioPortal SPARQL - http://sparql.bioontology.org, Accessed Nov 2013. (Cited on page 106.)

[197] NCBO BioPortal. BioPortal result - http://sparql.bioontology.org/?query=PREFIX+omv%3A+%3Chttp%3A%2F%2Fomv.ontoware.org%2F2005%2F05%2Fontology%23%3Econstruct%0D%0A%7B%3Chttp%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FOBI_0001705%3E+%3Fp+%3Fo%7D+where+%7B%3Chttp%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FOBI_0001705%3E+%3Fp+%3Fo%7D&csrfmiddlewaretoken=3aaacdd38fdfa7d174964853b92a1e38, Accessed Nov 2013. (Cited on page 106.)

[198] JA Blake, J Corradi, JT Eppig, DP Hill, JE Richardson, M Ringwald, et al. Creating the gene ontology resource: design and implementation., 2001. (Cited on page 107.)

[199] Ontotext. Linked Life Data - http://linkedlifedata.com/, Accessed Nov 2013. (Cited on page 107.)

[200] Richard Cyganiak and Chistian Bizer. Pubby-a linked data frontend for sparql endpoints - http://www4.wiwiss.fu-berlin.de/pubby/, Accessed Nov 2013. (Cited on page 107.)

[201] Bio2RDF Term search - http://bio2rdf.org/page/go:0032283, Accessed Nov 2013. (Cited on page 107.)

[202] IAO Term search - http://purl.obolibrary.org/obo/IAO_0000032, Accessed Nov 2013. (Cited on page 108.)

[203] Bio2RDF Term search - http://bio2rdf.org/page/iao:0000032, Accessed Nov 2013. (Cited on page 108.)

[204] Bio2RDF Term search - http://bio2rdf.org/taxon:9913, Accessed Nov 2013. (Cited on page 108.)

[205] OpenLink. Openlink's Virtuoso Faceted Browser term display - http://s4.semanticscience.org:16027/describe/?url=http%3A%2F%2Fbio2rdf.org%2Ftaxon%3A9913, Accessed Nov 2013. (Cited on page 108.)

[206] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. Hypertext transfer protocol–HTTP/1.1, 1999. (Cited on page 109.)

[207] OBO Foundry. OBO library - `http://obofoundry.org/`, Accessed Nov 2013. (Cited on page 112.)

[208] Christopher J Mungall. OBO2OWL pipeline - `https://github.com/cmungall/obo2owl`, Accessed Nov 2013. (Cited on page 112.)

[209] Christopher J Mungall. OBO flat file format 1.4 syntax and semantics [draft]. Technical report, Lawrence Berkeley National Laboratory. Available at `http://berkeleybop.org/~cjm/obo2owl/obo-syntax.html`. Accessed 5 Mar, 2012. (Cited on page 112.)

[210] Patrick Stickler. CBD-concise bounded description - `http://www.w3.org/Submission/CBD/`. *W3C Member Submission*, Accessed Nov 2013. (Cited on page 113.)

[211] OpenLink Software. SPARQL describe - `http://docs.openlinksw.com/virtuoso/rdfsparql.html#rdfsqlfromsparqldescribe`, June 2011. (Cited on page 113.)

[212] jQuery API documentation - `http://api.jquery.com`, Accessed Nov 2013. (Cited on page 114.)

[213] Jesse James Garrett et al. Ajax: A new approach to web applications - `http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications`, 2005. (Cited on page 114.)

[214] Fahim T Imam, Stephen D Larson, Anita Bandrowski, Jeffery S Grethe, Amarnath Gupta, Maryann E Martone, et al. Development and use of ontologies inside the neuroscience information framework: a practical approach. *Frontiers in genetics*, 3, 2012. (Cited on page 122.)

[215] Dan Brickley and Libby Miller. FOAF vocabulary specification 0.98 - `http://xmlns.com/foaf/spec/`. *Namespace document*, Accessed Nov 2013. (Cited on page 122.)

[216] Vaccine Ontology, *vaccination* - `http://purl.obolibrary.org/obo/VO_0000002`. (Cited on page 124.)

[217] The Ontology for Biomedical Investigations, *administering substance in vivo* - `http://purl.obolibrary.org/obo/OBI_0600007`. (Cited on page 124.)

[218] OMRE developers group. Ontology of Medically Relevant Entities (OMRE) - `http://code.google.com/p/ogms/wiki/OMRE`, December 2012. (Cited on page 124.)

[219] B J Stewart and P U Prabhu. Reports of sensorineural deafness after measles, mumps, and rubella immunisation. *Archives of Disease in Childhood*, 69(1):153–154, 1993. (Cited on page 124.)

[220] Madhok R Alcorn N, Saunders S. Benefit-risk assessment of leflunomide: an appraisal of leflunomide in rheumatoid arthritis 10 years after licensing. *Drug Safety*, 32(12):1123–34, 2009. (Cited on page 124.)

[221] Lidian L. A. Lecluse, Emmilia A. Dowlatshahi, C. E. Jacqueline M. Limpens, Menno A. de Rie, Jan D. Bos, and Phyllis I. Spuls. Etanercept: An overview of dermatologic adverse events. *Arch Dermatol*, 147(1):79–94, 2011. (Cited on page 124.)

[222] Angela A. M. C. Claessens, Eibert R. Heerdink, Jacques T. H. M. van Eijk, Cornelis B. H. W. Lamers, and Hubert G. M. Leufkens. Determinants of Headache in Lansoprazole Users in The Netherlands: Results from a Nested Case-Control Study. *Drug Safety*, 25(4), 2002. (Cited on page 124.)

[223] The Ontology of Medically Relevant Entities, *low blood pressure* - `http://purl.obolibrary.org/obo/ogms/OMRE_0000037`. (Cited on page 126.)

[224] Information Artifact Ontology, *is about* - `http://purl.obolibrary.org/obo/IAO_0000136`. (Cited on page 126.)

[225] The Adverse Event Reporting Ontology, *has component* - `http://purl.obolibrary.org/obo/AERO_0000125`. (Cited on pages 127 and 131.)

[226] The Adverse Event Reporting Ontology, *found to exhibit* - `http://purl.obolibrary.org/obo/AERO_0000088`. (Cited on page 127.)

[227] Information Artifact Ontology, *directive information entity* - `http://purl.obolibrary.org/obo/IAO_0000033`. (Cited on page 128.)

[228] Basic Formal Ontology, *realizable entity* - `http://www.ifomis.org/bfo/1.0/snap#RealizableEntity`. (Cited on page 128.)

[229] The Adverse Event Reporting Ontology, *level 1 of certainty of anaphylaxis according to Brighton* - `http://purl.obolibrary.org/obo/AERO_0000269`. (Cited on page 128.)

[230] The Ontology for General Medical Science, *clinical finding* - `http://purl.obolibrary.org/obo/OGMS_0000014`. (Cited on page 128.)

[231] Kevin P. High, Suzanne F. Bradley, Stefan Gravenstein, David R. Mehr, Vincent J. Quagliarello, Chesley Richards, and Thomas T. Yoshikawa. Clinical practice guideline for the evaluation of fever and infection in older adult residents of long-term care facilities: 2008 update by the infectious diseases society of america. *Clinical Infectious Diseases*, 48(2):149–171, 2009. (Cited on page 129.)

[232] Walter T. Hughes, Donald Armstrong, Gerald P. Bodey, Eric J. Bow, Arthur E. Brown, Thierry Calandra, Ronald Feld, Philip A. Pizzo, Kenneth V. I. Rolston, Jerry L. Shenep, and Lowell S. Young. 2002 guidelines for the use of antimicrobial agents in neutropenic patients with cancer. *Clinical Infectious Diseases*, 34(6):730–751, 2002. (Cited on page 129.)

[233] The Adverse Event Reporting Ontology, *chest tightness finding* - `http://purl.obolibrary.org/obo/AERO_0000356`. (Cited on page 129.)

[234] The Adverse Event Reporting Ontology, *measured hypotension finding* - `http://purl.obolibrary.org/obo/AERO_0000251`. (Cited on page 129.)

[235] C.J. Mungall, C. Torniai, G.V. Gkoutos, S.E. Lewis, and M.A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biology*, 13(1):R5, 2012. (Cited on page 129.)

[236] The Relation Ontology, *has part* - `http://www.obofoundry.org/ro/ro.owl#has_part`. (Cited on page 131.)

[237] World Health Organization. Severe falciparum malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 94:1–90, 2000. (Cited on page 132.)

[238] Information Artifact Ontology, *scalar measurement datum* - `http://purl.obolibrary.org/obo/IAO_0000032`. (Cited on page 132.)

[239] M. Krupka, K. Seydel, C.M. Feintuch, K. Yee, R. Kim, C.Y. Lin, R.B. Calder, C. Petersen, T. Taylor, and J. Daily. Mild Plasmodium falciparum malaria following an episode of severe malaria is associated with induction of the interferon pathway in Malawian children. *Infection and immunity*, 80(3):1150–1155, 2012. (Cited on pages 132 and 133.)

[240] Remi Gagnon, Marie Noel Primeau, Anne Des Roches, Chantal Lemire, Rhoda Kagan, Stuart Carr, Manale Ouakki, Mélanie Benoît, and Gaston De Serres. Safe vaccination of patients with egg allergy with an adjuvanted pandemic H1N1 vaccine. *Journal of Allergy and Clinical Immunology*, 126(2):317–323, 2010. (Cited on page 133.)

[241] National Institute of Allergy and Infectious Diseases/Food Allergy and Anaphylaxis Network anaphylaxis guideline - `http://www.niaid.nih.gov/topics/allergicDiseases/understanding/Pages/Anaphylaxis.aspx`, December 2012. (Cited on page 133.)

[242] Paul Shekelle, Martin P Eccles, Jeremy M Grimshaw, and Steven H Woolf. When should clinical guidelines be updated? *BMJ*, 323(7305):155–157, 7 2001. (Cited on page 133.)

[243] Taxiarchis Botsis, Michael D Nguyen, Emily Jane Woo, Marianthi Markatou, and Robert Ball. Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association: JAMIA*, 18(5):631–638, October 2011. (Cited on pages 135, 138, and 156.)

[244] Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA), agencies of the U.S. Department of Health and Human Services. Vaccine Adverse Event Reporting System (VAERS) - `http://vaers.hhs.gov/index.`, June Retrieved 2012. (Cited on page 135.)

[245] Report of Adverse Events Following Immunization form - `http://www.phac-aspc.gc.ca/im/pdf/raefi-dmcisi-eng.pdf`, December 2012. (Cited on page 135.)

[246] Taxiarchis Botsis, EmilyJane Woo, and Robert Ball. Application of information retrieval approaches to case classification in the vaccine adverse event reporting system. *Drug Safety*, 36(7):573–582, 2013. (Cited on pages 138, 139, 140, 141, 143, 144, 145, 146, and 149.)

[247] FuXi 1.0: A Python-based, bi-directional logical reasoningsystem - `http://code.google.com/p/fuxi/`, August 2013. (Cited on page 139.)

[248] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001. (Cited on page 144.)

[249] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Muller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77, 2011. (Cited on page 146.)

[250] Barbara A Slade, Laura Leidel, Claudia Vellozzi, Emily Jane Woo, Wei Hua, Andrea Sutherland, Hector S Izurieta, Robert Ball, Nancy Miller, M Miles Braun, et al. Postlicensure safety surveillance for quadrivalent human papillomavirus recombinant vaccine. *JAMA: the journal of the American Medical Association*, 302(7):750–757, 2009. (Cited on page 147.)

[251] WHO Global Advisory Committee on Vaccine Safety. Report of meeting held 12-13 June 2013 - `http://www.who.int/vaccine_safety/committee/reports/Jun_2013/en/index.html`, October Accessed Oct 2013. (Cited on page 148.)

[252] Pedro L Moro, Theresa Harrington, Tom Shimabukuro, Maria Cano, Oidda I Museru, David Menschik, and Karen Broder. Adverse events after Fluzone Intradermal vaccine reported to the Vaccine Adverse Event Reporting System (VAERS), 2011–2013. *Vaccine*, 2013. (Cited on page 148.)

[253] John M Kelso, Gina T Mootrey, and Theodore F Tsai. Anaphylaxis from yellow fever vaccine. *Journal of allergy and clinical immunology*, 103(4):698–701, 1999. (Cited on page 148.)

[254] Lauren DiMiceli, Vitali Pool, John M Kelso, Sean V Shadomy, John Iskander, and VAERS Team. Vaccination of yeast sensitive individuals: review of safety data in the US vaccine adverse event reporting system (VAERS). *Vaccine*, 24(6):703–707, 2006. (Cited on page 148.)

[255] Nicole P Lindsey, Betsy A Schroeder, Elaine R Miller, M Miles Braun, Alison F Hinckley, Nina Marano, Barbara A Slade, Elizabeth D Barnett, Gary W Brunette, Katherine Horan, et al. Adverse event reports following yellow fever vaccination. *Vaccine*, 26(48):6077–6082, 2008. (Cited on page 148.)

[256] John K Iskander, Elaine R Miller, and Robert T Chen. Vaccine adverse event reporting system (VAERS). *Pediatr Ann*, 33:599, 2004. (Cited on page 148.)

[257] Joel J Gagnier, Gunver Kienle, Douglas G Altman, David Moher, Harold Sox, and David Riley. The CARE guidelines: consensus-based clinical case report guideline development. *Journal of clinical epidemiology*, 2013. (Cited on page 148.)

[258] Anjan K Banerjee, Sally Okun, I Ralph Edwards, Paul Wicks, Meredith Y Smith, Stephen J Mayall, Bruno Flamion, Charles Cleeland, and Ethan Basch. Patient-reported outcome measures in safety event reporting: Prosper consortium guidance. *Drug Safety*, pages 1–21, 2013. (Cited on page 148.)

[259] Janice Minard, Suzanne M Dostaler, Jennifer G Olajos-Clow, Todd W Sands, Chris J Licskai, and M Diane Lougheed. Development and implementation of an electronic asthma record for primary care: Integrating guidelines into practice. *Journal of Asthma*, pages 1–29, 2013. (Cited on page 148.)

[260] David W Scheifele and Scott A Halperin. Immunization monitoring program, active: a model of active surveillance of vaccine safety. In *Seminars in pediatric infectious diseases*, volume 14, pages 213–219. Elsevier, 2003. (Cited on page 151.)

[261] Stella Veretnik, J Lynn Fink, and Philip E Bourne. Computational biology resources lack persistence and usability. *PLoS computational biology*, 4(7):e1000136, 2008. (Cited on page 152.)

[262] Jonathan D Wren and Alex Bateman. Databases, data tombs and dust in the wind. *Bioinformatics*, 24(19):2127–2128, 2008. (Cited on page 152.)

[263] Jonathan D Wren. URL decay in MEDLINEa 4-year follow-up study. *Bioinformatics*, 24(11):1381–1385, 2008. (Cited on page 152.)

[264] Inc. Dice Holdings. Sourceforge - `http://sourceforge.net`, Accessed Jan 2014. (Cited on page 152.)

[265] Google. Google code - `http://code.google.com`, Accessed Jan 2014. (Cited on page 152.)

[266] Mélanie Courtot and the OBO TWG. PURL Guide - `http://code.google.com/p/obo-foundry-operations-committee/wiki/PURLGuide`, Accessed Jan 2014. (Cited on page 153.)

[267] Mélanie Courtot and the OBO TWG. OBO PURL domain - `http://code.google.com/p/obo-foundry-operations-committee/wiki/OBOPURLDomain`, Accessed Jan 2014. (Cited on page 153.)

[268] Mélanie Courtot and the OBO TWG. Setting up Protege to work with OBO ontologies- `http://code.google.com/p/obo-foundry-operations-committee/wiki/SettingUpProtege`, Accessed Jan 2014. (Cited on page 153.)

[269] Mélanie Courtot and the OBO TWG. Policy for OBO namespace and associated PUR requests- http://code.google.com/p/obo-foundry-operations-committee/wiki/Policy_for_OBO_namespace_and_associated_PURL_requests, Accessed Jan 2014. (Cited on page 153.)

[270] Catia Pesquita, João D Ferreira, Francisco M Couto, and Mário J Silva. The epidemiology ontology: an ontology for the semantic annotation of epidemiological resources. *Journal of Biomedical Semantics*, 5(1):4, 2014. (Cited on page 153.)

[271] MSrv developers. MSrv - An ontology of medical surveillance- http://code.google.com/p/msrv/, Accessed Jan 2014. (Cited on page 153.)

[272] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, 2008. (Cited on page 156.)

Appendix A

Canadian Adverse Events Following Immunization Surveillance System (CAEFISS) sample data

Health Canada Santé Canada

Canada Vigilance
Summary of Reported Adverse Reactions

Report Runtime:	2011-09-02 - 05:27:16 PM
Initial Received Date:	1965-01-01 to 2011-03-31
Latest Received Date:	N/A
Total Number of Reports:	10 Report(s)

Brand Name/Active Ingredient: 'GARDASIL'
Search Date Criteria: 1965-01-01 to 2011-03-31
Reaction Term(s): All/Tous
Serious report?: Both
Feature of Report: All
Type of Report: All
Source of Report: All
Gender: All
Report Outcome: All
Age: All

CAVEAT: This summary is based on information from adverse reaction reports submitted by health professionals and laypersons either directly to Health Canada or via market authorization holders. Each report represents the suspicion, opinion or observation of the individual reporter. The Canada Vigilance Program is a spontaneous reporting system that is suitable to detect signals of potential health product safety issues during the post-market period. The data has been collected primarily by a spontaneous surveillance system in which adverse reactions to health products are reported on a voluntary basis. Under reporting of adverse reactions is seen with both voluntary and mandatory spontaneous surveillance systems. Accumulated case reports should not be used as a basis for determining the incidence of a reaction or estimating risk for a particular product as neither the total number of reactions occurring, nor the number of patients exposed to the health product is known. Because of the multiple factors that influence reporting, quantitative comparisons of health product safety cannot be made from the data. Some of these factors include the length of time a drug is marketed, the market share, size and sophistication of the sales force, publicity about an adverse reaction and regulatory actions. In some cases, the reported clinical data is incomplete and there is not certainty that these health products caused the reported reactions. A given reaction may be due to an underlying disease process or to another coincidental factor. This information is provided with the understanding that the data will be appropriately referenced and used in conjunction with this caveat statement.

Canada

Canada Vigilance
Summary of Reported Adverse Reactions

Report Runtime:	2011-09-02 - 05:27:16 PM
Initial Received Date:	1965-01-01 to 2011-03-31
Latest Received Date:	N/A
Total Number of Reports:	10 Report(s)

Report Information ***AER = Adverse Reaction Report*

Adverse Reaction Report Number	Latest AER** Version Number	Initial Received Date	Latest Received Date	Source of Report	Market Authorization Holder AER Number	Feature of Report	Type of Report	Reporter Type
000358593	0	2010-12-24	2010-12-24	MAH	2010004848	Adverse Reaction	Spontaneous	Health Professional

Serious report?	Death:	Disability:	Congenital Anomaly:
Yes	Yes		
	Life Threatening:	**Hospitalization:** Yes	**Other Medically Important Conditions:** Yes

Patient Information

Age	Gender	Height	Weight	Report Outcome
14 Years	Female			Death

Link / Duplicate Report Information

Record Type	Link AER** Number
No duplicate or linked report.	

Product Information

Product Description	Health Product Role	Dosage Form	Route of Administration	Dose	Frequency	Therapy Duration
GARDASIL	Suspect		Unknown			1.0 Day(s)
GARDASIL	Suspect	SUSPENSION INTRAMUSCULAR	Subcutaneous			1.0 Day(s)
INFLUENZA VACCINE	Concomitant	NOT SPECIFIED	Unknown			
YASMIN 21	Suspect	TABLET	Unknown			61.0 Day(s)

Adverse Reaction Term Information

Adverse Reaction Term(s)	MedDRA Version	Reaction Duration
Abasia	v.14.0	
Asthenia	v.14.0	
Basilar migraine	v.14.0	
Blood glucose increased	v.14.0	
Cardiac arrest	v.14.0	

Adverse Reaction Term(s)	MedDRA Version	Reaction Duration
Confusional state	v.14.0	
Dizziness postural	v.14.0	
Drowning	v.14.0	
Loss of consciousness	v.14.0	
Nausea	v.14.0	
Syncope	v.14.0	
Vomiting	v.14.0	

Canada Vigilance
Summary of Reported Adverse Reactions

Report Runtime: 2011-09-02 - 05:27:16 PM
Initial Received Date: 1965-01-01 to 2011-03-31
Latest Received Date: N/A
Total Number of Reports: 10 Report(s)

Report Information ***AER = Adverse Reaction Report*

Adverse Reaction Report Number	Latest AER** Version Number	Initial Received Date	Latest Received Date	Source of Report	Market Authorization Holder AER Number	Feature of Report	Type of Report	Reporter Type
000359862	0	2011-01-17	2011-01-17	Community		Adverse Reaction	Spontaneous	Physician

Serious report?	Death:		Disability:		Congenital Anomaly:	
Yes	Life Threatening:	Yes	Hospitalization:	Yes	Other Medically Important Conditions:	

Patient Information

Age	Gender	Height	Weight	Report Outcome
14 Years	Female	158 Centimetres	80 Kilograms	Unknown

Link / Duplicate Report Information

Record Type	Link AER** Number
Duplicate	000360728

Product Information

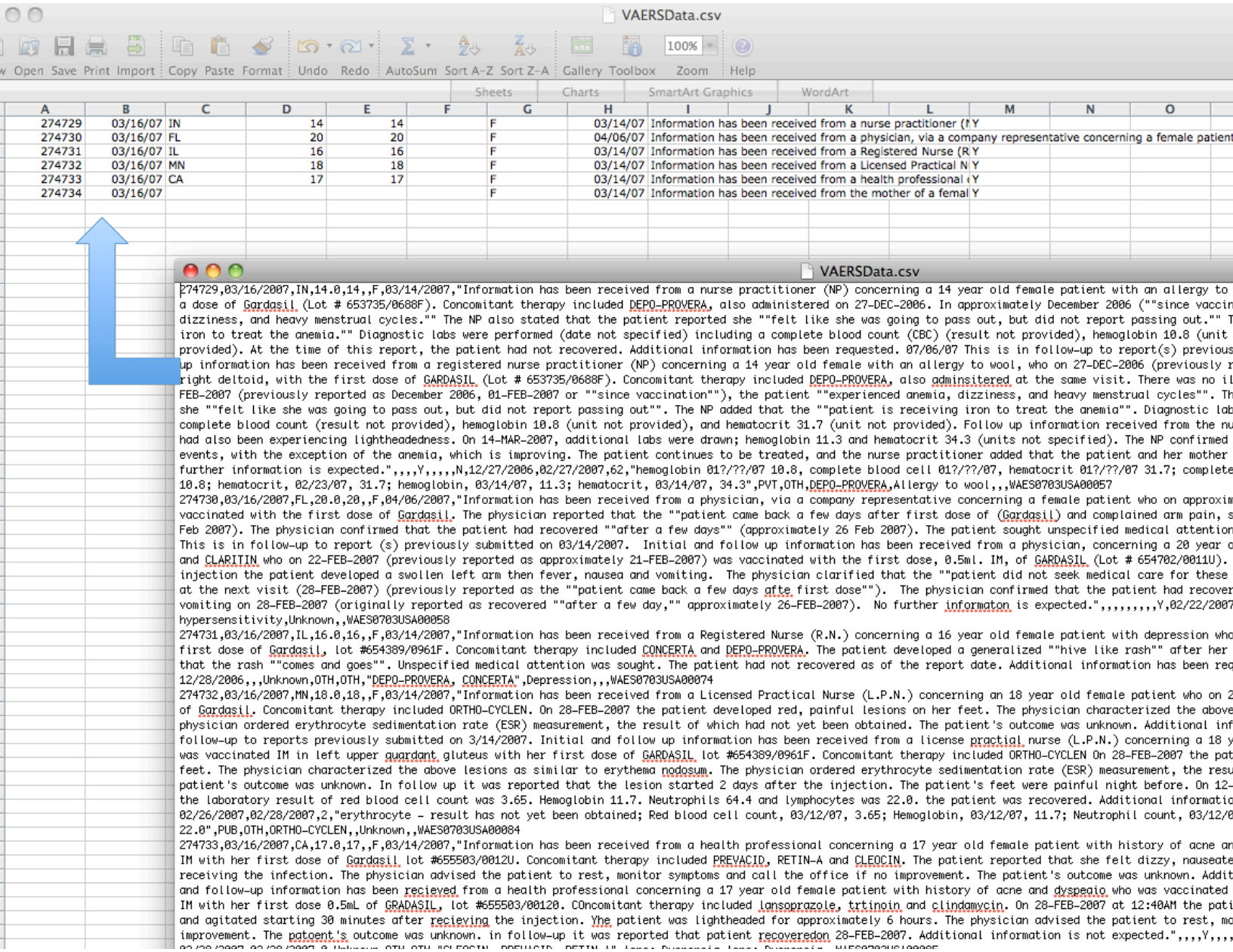
Product Description	Health Product Role	Dosage Form	Route of Administration	Dose	Frequency	Therapy Duration
GARDASIL	Suspect	SUSPENSION INTRAMUSCULAR	Unknown	1.0 Dosage forms	Once	

Adverse Reaction Term Information

Adverse Reaction Term(s)	MedDRA Version	Reaction Duration
Nervous system disorder	v.14.0	
Ventricular fibrillation	v.14.0	

Appendix B

Vaccine Adverse Event Reporting System (VAERS) sample data


VAERSData.csv
Open Save Print Import Copy Paste Format Undo Redo AutoSum Sort A-Z Sort Z-A Gallery Toolbox 100% Zoom Help
Sheets Charts SmartArt Graphics WordArt
A B C D E F G H I J K L M N O
274729 03/16/07 IN 14 14 F 03/14/07 Information has been received from a nurse practitioner (N Y
274730 03/16/07 FL 20 20 F 04/06/07 Information has been received from a physician, via a company representative concerning a female patient
274731 03/16/07 IL 16 16 F 03/14/07 Information has been received from a Registered Nurse (R Y
274732 03/16/07 MN 18 18 F 03/14/07 Information has been received from a Licensed Practical N Y
274733 03/16/07 CA 17 17 F 03/14/07 Information has been received from a health professional Y
274734 03/16/07 F 03/14/07 Information has been received from the mother of a femal Y
VAERSData.csv
274729,03/16/2007,IN,14.0,14,,F,03/14/2007,"Information has been received from a nurse practitioner (NP) concerning a 14 year old female patient with an allergy to
a dose of Gardasil (Lot # 653735/0688F). Concomitant therapy included DEPO-PROVERA, also administered on 27-DEC-2006. In approximately December 2006 (""since vaccin
dizziness, and heavy menstrual cycles."" The NP also stated that the patient reported she ""felt like she was going to pass out, but did not report passing out."" T
iron to treat the anemia."" Diagnostic labs were performed (date not specified) including a complete blood count (CBC) (result not provided), hemoglobin 10.8 (unit
provided). At the time of this report, the patient had not recovered. Additional information has been requested. 07/06/07 This is in follow-up to report(s) previous
up information has been received from a registered nurse practitioner (NP) concerning a 14 year old female with an allergy to wool, who on 27-DEC-2006 (previously r
right deltoid, with the first dose of GARDASIL (Lot # 653735/0688F). Concomitant therapy included DEPO-PROVERA, also adminsitered at the same visit. There was no il
FEB-2007 (previously reported as December 2006, 01-FEB-2007 or ""since vaccination""), the patient ""experienced anemia, dizziness, and heavy menstrual cycles"". Th
she ""felt like she was going to pass out, but did not report passing out"". The NP added that the ""patient is receiving iron to treat the anemia"". Diagnostic lab
complete blood count (result not provided), hemoglobin 10.8 (unit not provided), and hematocrit 31.7 (unit not provided). Follow up information received from the nu
had also been experiencing lightheadedness. On 14-MAR-2007, additional labs were drawn; hemoglobin 11.3 and hematocrit 34.3 (units not specified). The NP confirmed
events, with the exception of the anemia, which is improving. The patient continues to be treated, and the nurse practitioner added that the patient and her mother
further information is expected.",,,,Y,,,,,N,12/27/2006,02/27/2007,62,"hemoglobin 01?/??/07 10.8, complete blood cell 01?/??/07, hematocrit 01?/??/07 31.7; complete
10.8; hematocrit, 02/23/07, 31.7; hemoglobin, 03/14/07, 11.3; hematocrit, 03/14/07, 34.3",PVT,OTH,DEPO-PROVERA,Allergy to wool,,,WAES0703USA00057
274730,03/16/2007,FL,20.0,20,,F,04/06/2007,"Information has been received from a physician, via a company representative concerning a female patient who on approxim
vaccinated with the first dose of Gardasil. The physician reported that the ""patient came back a few days after first dose of (Gardasil) and complained arm pain, s
Feb 2007). The physician confirmed that the patient had recovered ""after a few days"" (approximately 26 Feb 2007). The patient sought unspecified medical attention
This is in follow-up to report (s) previously submitted on 03/14/2007. Initial and follow up information has been received from a physician, concerning a 20 year o
and CLARITIN who on 22-FEB-2007 (previously reported as approximately 21-FEB-2007) was vaccinated with the first dose, 0.5ml. IM, of GARDASIL (Lot # 654702/0011U).
injection the patient developed a swollen left arm then fever, nausea and vomiting. The physician clarified that the ""patient did not seek medical care for these
at the next visit (28-FEB-2007) (previously reported as the ""patient came back a few days afte first dose""). The physician confirmed that the patient had recover
vomiting on 28-FEB-2007 (originally reported as recovered ""after a few day,"" approximately 26-FEB-2007). No further informaton is expected.",,,,,,,,,Y,02/22/2007
hypersensitivity,Unknown,,WAES0703USA00058
274731,03/16/2007,IL,16.0,16,,F,03/14/2007,"Information has been received from a Registered Nurse (R.N.) concerning a 16 year old female patient with depression who
first dose of Gardasil, lot #654389/0961F. Concomitant therapy included CONCERTA and DEPO-PROVERA. The patient developed a generalized ""hive like rash"" after her
that the rash ""comes and goes"". Unspecified medical attention was sought. The patient had not recovered as of the report date. Additional information has been req
12/28/2006,,,Unknown,OTH,OTH,"DEPO-PROVERA, CONCERTA",Depression,,,WAES0703USA00074
274732,03/16/2007,MN,18.0,18,,F,03/14/2007,"Information has been received from a Licensed Practical Nurse (L.P.N.) concerning an 18 year old female patient who on 2
of Gardasil. Concomitant therapy included ORTHO-CYCLEN. On 28-FEB-2007 the patient developed red, painful lesions on her feet. The physician characterized the above
physician ordered erythrocyte sedimentation rate (ESR) measurement, the result of which had not yet been obtained. The patient's outcome was unknown. Additional inf
follow-up to reports previously submitted on 3/14/2007. Initial and follow up information has been received from a license practial nurse (L.P.N.) concerning a 18 y
was vaccinated IM in left upper quardant gluteus with her first dose of GARDASIL lot #654389/0961F. Concomitant therapy included ORTHO-CYCLEN On 28-FEB-2007 the pat
feet. The physician characterized the above lesions as similar to erythema nodosum. The physician ordered erythrocyte sedimentation rate (ESR) measurement, the resu
patient's outcome was unknown. In follow up it was reported that the lesion started 2 days after the injection. The patient's feet were painful night before. On 12-
the laboratory result of red blood cell count was 3.65. Hemoglobin 11.7. Neutrophils 64.4 and lymphocytes was 22.0. the patient was recovered. Additional informatio
02/26/2007,02/28/2007,2,"erythrocyte - result has not yet been obtained; Red blood cell count, 03/12/07, 3.65; Hemoglobin, 03/12/07, 11.7; Neutrophil count, 03/12/0
22.0",PUB,OTH,ORTHO-CYCLEN,,Unknown,,WAES0703USA00084
274733,03/16/2007,CA,17.0,17,,F,03/14/2007,"Information has been received from a health professional concerning a 17 year old female patient with history of acne an
IM with her first dose of Gardasil lot #655503/0012U. Concomitant therapy included PREVACID, RETIN-A and CLEOCIN. The patient reported that she felt dizzy, nauseate
receiving the infection. The physician advised the patient to rest, monitor symptoms and call the office if no improvement. The patient's outcome was unknown. Addit
and follow-up information has been recieved from a health professional concerning a 17 year old female patient with history of acne and dyspeaia who was vaccinated
IM with her first dose 0.5mL of GRADASIL, lot #655503/00120. COncomitant therapy included lansoprazole, trtinoin and clindamycin. On 28-FEB-2007 at 12:40AM the pati
and agitated starting 30 minutes after recieving the injection. Yhe patient was lightheaded for approximately 6 hours. The physician advised the patient to rest, mo
improvement. The patoent's outcome was unknown. in follow-up it was reported that patient recoveredon 28-FEB-2007. Additional information is not expected.",,,,Y,,,,

Appendix C

List of OBO Foundry principles (as of November 2013)

Principle ID	Description
Accepted	
FP 001 open	The ontology must be open and available to be used by all without any constraint other than (a) its origin must be acknowledged and (b) it is not to be altered and subsequently redistributed under the original name or with the same identifiers. The OBO ontologies are for sharing and are resources for the entire community. For this reason, they must be available to all without any constraint or license on their use or redistribution. However, it is proper that their original source is always credited and that after any external alterations, they must never be redistributed under the same name or with the same identifiers.
FP 002 format	The ontology is in, or can be expressed in, a common shared syntax. This may be either the OBO syntax, extensions of this syntax, or OWL. The reason for this is that the same tools can then be usefully applied. This facilitates shared software implementations. This criterion is not met in all of the ontologies currently listed, but we are working with the ontology developers to have them available in a common OBO syntax.
FP 003 URIs	The ontologies possess a unique identifier space within the OBO Foundry. The source of a term (i.e. class) from any ontology can be immediately identified by the prefix of the identifier of each term. It is, therefore, important that this prefix be unique.

Principle ID	**Description**
FP 004 versioning	The ontology provider has procedures for identifying distinct successive versions.
FP 005 delineated content	The ontology has a clearly specified and clearly delineated content. The ontology must be orthogonal to other ontologies already lodged within OBO. The major reason for this principle is to allow two different ontologies, for example anatomy and process, to be combined through additional relationships. These relationships could then be used to constrain when terms could be jointly applied to describe complementary (but distinguishable) perspectives on the same biological or medical entity. As a corollary to this, we would strive for community acceptance of a single ontology for one domain, rather than encouraging rivalry between ontologies.
FP 006 textual definition	The ontologies include textual definitions for all terms. Many biological and medical terms may be ambiguous, so terms should be defined so that their precise meaning within the context of a particular ontology is clear to a human reader.
FP 007 relations	The ontology uses relations which are unambiguously defined following the pattern of definitions laid down in the OBO Relation Ontology.
FP 008 documented	The ontology is well documented.
FP 009 users	The ontology has a plurality of independent users.
FP 010 collaboration	The ontology will be developed collaboratively with other OBO Foundry members.
FP 011 locus of authority	There should be a single person who is responsible for the ontology, for ensuring continued maintenance in light of scientific advance and prompt response to user feedback, Contact information for this person should be provided on the ontology website, and listed in the OBO Library Metadata File.
FP 012 naming conventions	The ontology follows the OBO set of naming conventions.

Principle ID	**Description**
FP 016 maintenance	OBO is an open community and, by joining the initiative, the authors of an ontology commit to its maintenance in light of scientific advance and to working with other members to ensure the improvement of these principles over time.
Under discussion	
FP 013 genus differentia	All definitions of the genus-differentia form, utilizing (some) cross-products.
FP 014 BFO	Ontologies should be conceivable as the result of populating downwards from some fragment of BFO2.0.
FP 015 Single inheritance	Single asserted is_a inheritance (= each ontology should be conceived as consisting of a core of asserted single inheritance links, with further is_a relations inferred).
FP 017 instantiability	All the types represented by the terms in the ontology should be instantiable.
FP 018 orthogonality	For each domain there should be convergence upon a single ontology that is recommended for use by those who wish to become involved with the Foundry initiative.
FP 019 content	The ontology must be a faithful representation of the domain and fit for the stated purpose.

Appendix D

SPARQL query for FluMist vaccine

```
DEFINE sql:describe-mode "CBD"
describe <http://purl.obolibrary.org/obo/VO_0000044>
FROM <http://purl.obolibrary.org/obo/merged/VO>

================================================================

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix owl: <http://www.w3.org/2002/07/owl#>

select *
from <http://purl.obolibrary.org/obo/merged/VO>
where {
?nodeID owl:annotatedSource <http://purl.obolibrary.org/obo/
VO_0000044>.
#?nodeID rdf:type owl:Annotation.
?nodeID owl:annotatedProperty ?annotatedProperty.
?nodeID owl:annotatedTarget ?annotatedTarget.
?nodeID ?aaProperty ?aaPropertyTarget.
OPTIONAL {?annotatedProperty rdfs:label ?annotatedPropertyLabel}.
OPTIONAL {?aaProperty rdfs:label ?aaPropertyLabel}.
FILTER (isLiteral(?annotatedTarget)).
FILTER (not (?aaProperty in(owl:annotatedSource, rdf:type,
owl:annotatedProperty, owl:annotatedTarget)))
}
================================================================

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?ref ?refp ?label  ?o
FROM <http://purl.obolibrary.org/obo/merged/VO>
WHERE {
     ?ref ?refp ?o.
     FILTER (?refp IN (owl:equivalentClass, rdfs:subClassOf)).
     OPTIONAL {?ref rdfs:label ?label}.
     {
          {
               SELECT ?s ?o
               FROM <http://purl.obolibrary.org/obo/merged/VO>

               WHERE {
                    ?o ?p ?s .
                    FILTER (?p IN (rdf:first, rdf:rest,
owl:intersectionOf, owl:unionOf, owl:someValuesFrom, owl:hasValue,
owl:allValuesFrom, owl:complementOf, owl:inverseOf, owl:onClass,
```

```
owl:onProperty))
                }
          }
          OPTION (TRANSITIVE, t_in(?s), t_out(?o), t_step(?s) as ?
link).
          FILTER (?s= <http://purl.obolibrary.org/obo/VO_0000044>)
     }
}

ORDER BY ?label
================================================================

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?s ?o ?sc
FROM <http://purl.obolibrary.org/obo/merged/VO>
WHERE {
{
?s rdfs:subClassOf <http://purl.obolibrary.org/obo/VO_0000044> .
FILTER (isIRI(?s)).
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc rdfs:subClassOf ?s}
}
UNION
{
?s owl:equivalentClass ?s1 .
FILTER (isIRI(?s)).
?s1 owl:intersectionOf ?s2 .
?s2 rdf:first <http://purl.obolibrary.org/obo/VO_0000044> .
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc rdfs:subClassOf ?s}
}
UNION
{
?s rdfs:subClassOf <http://purl.obolibrary.org/obo/VO_0000044> .
FILTER (isIRI(?s)).
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc owl:equivalentClass ?s1 .
?s1 owl:intersectionOf ?s2 .
?s2 rdf:first ?s}
}
UNION
{
?s owl:equivalentClass ?s1 .
FILTER (isIRI(?s)).
```

```
?s1 owl:intersectionOf ?s2 .
?s2 rdf:first <http://purl.obolibrary.org/obo/VO_0000044> .
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc owl:equivalentClass ?s3 .
?s3 owl:intersectionOf ?s4 .
?s4 rdf:first ?s}
}
}
```

```
==========================================================

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix owl: <http://www.w3.org/2002/07/owl#>

SELECT ?path ?link ?label
FROM <http://purl.obolibrary.org/obo/merged/VO>
WHERE
{
{
SELECT ?s ?o ?label
WHERE
{{
?s rdfs:subClassOf ?o .
FILTER (isURI(?o)).
OPTIONAL {?o rdfs:label ?label}
}
UNION
{
?s owl:equivalentClass ?s1 .
?s1 owl:intersectionOf ?s2 .
?s2 rdf:first ?o  .
FILTER (isURI(?o))
OPTIONAL {?o rdfs:label ?label}
}
}
} OPTION (TRANSITIVE, t_in(?s), t_out(?o), t_step (?s) as ?link,
t_step ('path_id') as ?path).
FILTER (isIRI(?o)).
FILTER (?s= <http://purl.obolibrary.org/obo/VO_0000044>)
}
==========================================================

prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix owl: <http://www.w3.org/2002/07/owl#>

SELECT ?s ?label
```

```
FROM <http://purl.obolibrary.org/obo/merged/VO>
WHERE
{
     ?s rdf:type <http://purl.obolibrary.org/obo/VO_0000044> .
     ?s rdfs:label ?label
}
================================================================

SELECT distinct ?g
WHERE{
graph ?g
{
<http://purl.obolibrary.org/obo/VO_0000044> ?p ?o
}
}

================================================================

SELECT *
FROM <http://purl.obolibrary.org/obo/merged/VO>
WHERE { ?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?o.
FILTER (?s in(<http://null>, <http://purl.obolibrary.org/obo/
VO_0000044>, <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>,
<http://www.w3.org/2002/07/owl#Class>, <http://www.w3.org/2000/01/
rdf-schema#label>, <http://www.w3.org/2000/01/rdf-schema#comment>,
<http://www.w3.org/2000/01/rdf-schema#subClassOf>, <http://
purl.obolibrary.org/obo/VO_0000642>, <http://www.w3.org/2000/01/
rdf-schema#seeAlso>, <http://purl.obolibrary.org/obo/IAO_0000117>,
<http://www.w3.org/2002/07/owl#Restriction>, <http://www.w3.org/
2002/07/owl#onProperty>, <http://purl.obolibrary.org/obo/
VO_0003355>, <http://www.w3.org/2002/07/owl#someValuesFrom>,
<http://purl.obolibrary.org/obo/NCBITaxon_197911>, <http://
purl.obolibrary.org/obo/bearer_of>, <http://purl.obolibrary.org/
obo/VO_0000812>, <http://purl.obolibrary.org/obo/OBI_0000304>,
<http://www.w3.org/2002/07/owl#hasValue>, <http://
purl.obolibrary.org/obo/VO_0000694>, <http://purl.obolibrary.org/
obo/VO_0001243>, <http://purl.obolibrary.org/obo/NCBITaxon_9606>,
<http://purl.obolibrary.org/obo/VO_0000864>, <http://
purl.obolibrary.org/obo/NCBITaxon_197912>, <http://
purl.obolibrary.org/obo/VO_0000547>, <http://purl.obolibrary.org/
obo/VO_0000343>, <http://purl.obolibrary.org/obo/BFO_0000086>,
<http://purl.obolibrary.org/obo/VO_0001015>, <http://
purl.obolibrary.org/obo/VO_0000531>, <http://purl.obolibrary.org/
obo/CHEBI_7507>))
}

================================================================
```

```
SELECT *
FROM <http://purl.obolibrary.org/obo/merged/VO>
WHERE { ?s <http://www.w3.org/2000/01/rdf-schema#label> ?o.
FILTER (?s in(<http://null>, <http://purl.obolibrary.org/obo/
VO_0000044>, <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>,
<http://www.w3.org/2002/07/owl#Class>, <http://www.w3.org/2000/01/
rdf-schema#label>, <http://www.w3.org/2000/01/rdf-schema#comment>,
<http://www.w3.org/2000/01/rdf-schema#subClassOf>, <http://
purl.obolibrary.org/obo/VO_0000642>, <http://www.w3.org/2000/01/
rdf-schema#seeAlso>, <http://purl.obolibrary.org/obo/IAO_0000117>,
<http://www.w3.org/2002/07/owl#Restriction>, <http://www.w3.org/
2002/07/owl#onProperty>, <http://purl.obolibrary.org/obo/
VO_0003355>, <http://www.w3.org/2002/07/owl#someValuesFrom>,
<http://purl.obolibrary.org/obo/NCBITaxon_197911>, <http://
purl.obolibrary.org/obo/bearer_of>, <http://purl.obolibrary.org/
obo/VO_0000812>, <http://purl.obolibrary.org/obo/OBI_0000304>,
<http://www.w3.org/2002/07/owl#hasValue>, <http://
purl.obolibrary.org/obo/VO_0000694>, <http://purl.obolibrary.org/
obo/VO_0001243>, <http://purl.obolibrary.org/obo/NCBITaxon_9606>,
<http://purl.obolibrary.org/obo/VO_0000864>, <http://
purl.obolibrary.org/obo/NCBITaxon_197912>, <http://
purl.obolibrary.org/obo/VO_0000547>, <http://purl.obolibrary.org/
obo/VO_0000343>, <http://purl.obolibrary.org/obo/BFO_0000086>,
<http://purl.obolibrary.org/obo/VO_0001015>, <http://
purl.obolibrary.org/obo/VO_0000531>, <http://purl.obolibrary.org/
obo/CHEBI_7507>))
}

================================================================

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?s ?o ?sc
FROM <http://purl.obolibrary.org/obo/merged/VO>

WHERE {
{
?s rdfs:subClassOf <http://purl.obolibrary.org/obo/VO_0000642> .
FILTER (isIRI(?s)).
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc rdfs:subClassOf ?s}
}
UNION
{
?s owl:equivalentClass ?s1 .
```

```
FILTER (isIRI(?s)).
?s1 owl:intersectionOf ?s2 .
?s2 rdf:first <http://purl.obolibrary.org/obo/VO_0000642> .
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc rdfs:subClassOf ?s}
}
UNION
{
?s rdfs:subClassOf <http://purl.obolibrary.org/obo/VO_0000642> .
FILTER (isIRI(?s)).
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc owl:equivalentClass ?s1 .
?s1 owl:intersectionOf ?s2 .
?s2 rdf:first ?s}
}
UNION
{
?s owl:equivalentClass ?s1 .
FILTER (isIRI(?s)).
?s1 owl:intersectionOf ?s2 .
?s2 rdf:first <http://purl.obolibrary.org/obo/VO_0000642> .
OPTIONAL {?s rdfs:label ?o} .
OPTIONAL {?sc owl:equivalentClass ?s3 .
?s3 owl:intersectionOf ?s4 .
?s4 rdf:first ?s}
}
}
=================================================================
```

Appendix E

List of IAO annotation properties used as common metadata set

Label	Definition	Cardinality
editor preferred term	The concise, meaningful, and human-friendly name for a class or property preferred by the ontology developers. (US-English)	1:1
definition	The official definition, explaining the meaning of a class or property. Shall be Aristotelian, formalized and normalized. Can be augmented with colloquial definitions.	1:1
definition editor	Name of editor entering the definition in the file. The definition editor is a point of contact for information regarding the term. The definition editor may be, but is not always, the author of the definition, which may have been worked upon by several people.	1:n
definition source	formal citation, e.g., identifier in external database to indicate / attribute source(s) for the definition. Free text indicates attributes source(s) for the definition. EXAMPLE: Author Name, URI, MeSH Term C04, PUBMED ID, Wiki URI on 31.01.2007	1:n

Label	Definition	Cardinality
curation status specification	The curation status of a class or property. The allowed values must come from this enumerated list of predefined terms: • placeholder: This isn't a class that the ontology will keep - it's a placeholder for edits that are underway. The class name should start with an underscore • uncurated: Nothing done yet beyond assigning a unique class ID and proposing a preferred term • metadata incomplete: Class is being worked on; however, the metadata (including definition) are not complete or sufficiently clear to the editors. • metadata complete: Class has all its metadata, but is either not guaranteed to be in its final location in the asserted IS_A hierarchy or refers to another class that is not complete. • pending final vetting: All definitions, placement in the asserted IS_A hierarchy and required minimal metadata are complete. The class is awaiting a final review by someone other than the definition editor. • ready for release: Class has undergone final review, is ready for use, and will be included in the next release. Any class lacking "ready_for_release" should be considered likely to change place in hierarchy, have its definition refined, or be obsoleted in the next release. Those classes deemed "ready_for_release" will also derived from a chain of ancestor classes that are also "ready_for_release."	1:1

Label	Definition	Cardinality
example of usage	A phrase describing how a class name should be used. May also include other kinds of examples that facilitate immediate understanding of a class semantics, such as widely known prototypical subclasses or instances of the class. Although essential for high level terms, examples for low level terms (e.g., Affymetrix HU133 array) are not	0:n
alternative term	An alternative name for a class or property which means the same thing as the preferred name (semantically equivalent)	0:n
editor note	A note containing points under consideration for further term development that may be included in released versions of the ontology. It should contain nothing embarrassing and something potentially useful for end users to understand the ontology. Editor notes should include the date of edit (YYYYMMDD) and the author.	0:n
curator note	An administrative note intended for the curator of the ontology. It will not be included in the released versions of the ontology, so it should contain nothing necessary for end users to understand the ontology. Curator notes should include the date of edit (YYYY/MM/DD) and the author.	0:n

Label	Definition	Cardinality
obsolescence reason specification	The obsolescence reason of a class or property. The allowed values must come from this enumerated list of predefined terms: • failed exploratory term: The term was used in an attempt to structure part of the ontology but in retrospect failed to do a good job • terms merged: An editor note should explain what were the merged terms and the reason for the merge. • term split: This is to be used when a term has been split in two or more new terms. An editor note should indicate the reason for the split and indicate the URIs of the new terms created. • placeholder removed: This is to be used when the original term has been replaced by a term imported from an other ontology. An editor note should indicate what is the URI of the new term to use. • term imported: This is to be used when the original term has been replaced by a term imported from an other ontology. An editor note should indicate what is the URI of the new term to use.	1:1
OBO foundry unique label	An alternative name for a class or property which is unique across the OBO Foundry.	1:1

Appendix F

The anaphylactic reaction Standardised MedDRA Query

2.7 *Anaphylactic reaction (SMQ)*

(Production Release November 2005)

2.7.1 Definition

- An acute systemic reaction characterized by pruritus, generalized flush, urticaria, respiratory distress and vascular collapse
- Occurs in a previously sensitized person upon re-exposure to the sensitizing antigen
- Other signs and symptoms: agitation, palpitation, parasthesias, wheezing, angioedema, coughing, sneezing and difficulty breathing due to laryngeal spasm or bronchospasm
  - Less frequent clinical presentations: seizures, vomiting, abdominal cramps and incontinence

2.7.2 Inclusion/Exclusion Criteria

- Included:
  - Any terms, at the PT level, representing events which may be noted during anaphylaxis
  - In a spreadsheet format, the testing pharmaceutical company's list and the testing regulator's list were positioned alongside the MedDRA SSC list for anaphylaxis, and this three-column table was then systematically reviewed top-down. Unanimous agreement for/against inclusion of each term was achieved by the group
- Excluded:
  - Terms for signs and symptoms that do not fall within the three defined categories (Upper Airway/Respiratory, Angioedema/Urticaria/Pruritus/Flush, and Cardiovascular/Hypotension) in the broad search are excluded.

NOTE: There are two SMQs related to anaphylaxis: *Anaphylactic reaction (SMQ)* and *Anaphylactic/anaphylactoid shock conditions (SMQ)*. The two SMQs have different focuses. *Anaphylactic/anaphylactoid shock (SMQ)* is specific for more severe anaphylactic manifestations, i.e. those that result in shock, and not less severe ones such as rash. *Anaphylactic reaction (SMQ)* widens the search beyond shock conditions by including such terms as PT *Type I hypersensitivity*.

2.7.3 Algorithm

The SMQ *Anaphylactic reaction* consists of three parts:

- A **narrow search** containing PTs that represent core anaphylactic reaction terms;

- A **broad search** that contains additional terms that are added to those included in the narrow search. These additional terms are signs and symptoms possibly indicative of anaphylactic reaction;
- An **algorithmic approach** which combines a number of anaphylactic reaction symptoms in order to increase specificity. A case must include either:
  - A narrow term or a term from Category A;
  - A term from Category B - (Upper Airway/Respiratory) AND a term from Category C - (Angioedema/Urticaria/Pruritus/Flush);
  - A term from Category D - (Cardiovascular/Hypotension) AND [a term from Category B - (Upper Airway/Respiratory) OR a term from Category C - (Angioedema/Urticaria/ Pruritus/Flush)]

2.7.4 Notes on Implementation and/or Expectation of Query Results

In addition to narrow and broad searches, *Anaphylactic reaction (SMQ)* is an algorithmic SMQ. The algorithm is a combination of broad search terms among various categories to further refine the identification of cases of interest. The algorithm can be implemented in a post-retrieval process as noted below:

- First, retrieve relevant cases by applying the SMQ query as a narrow/broad SMQ (see section 1.5.2.1).
- Post-retrieval process, software applies the algorithmic combination to screen the cases retrieved above. For small data sets of retrieved cases, the algorithm may be applied on manual review of cases. The algorithm for *Anaphylactic reaction (SMQ)* is A or (B and C) or (D and (B or C)). Cases filtered by the algorithm can be listed for output.

2.7.5 List of References for *Anaphylactic reaction (SMQ)*

- The Merck Manual. 15th edition. Merck, Sharp & Dohme Research Laboratories. (1987): 306-7

Appendix G

The list of significant MedDRA terms based on contingency tables test

Supplementary material

Appendix 1: MedDRA terms with a chi-square value over 3.841

MedDRA term	Chi-square	P-value
Hypersensitivity	1578.605353	0
Dyspnoea	553.3557	2.34E-122
Throat tightness	551.5865009	5.69E-122
Pruritus	297.906177	9.42E-67
Chest discomfort	296.2635345	2.15E-66
Pharyngeal oedema	251.7630256	1.07E-56
Urticaria	231.0682725	3.49E-52
Wheezing	205.1667372	1.56E-46
Swelling face	203.0038003	4.62E-46
Anaphylactic reaction	198.3924991	4.68E-45
Oedema	181.4914781	2.29E-41
Swelling	179.028501	7.90E-41
Lip swelling	177.3909311	1.80E-40
Discomfort	160.1597406	1.04E-36
Swollen tongue	157.5517954	3.88E-36
Throat irritation	154.4938506	1.81E-35
Eye swelling	141.3551256	1.35E-32
Tic	122.0267653	2.28E-28
Dysphagia	83.93452989	5.11E-20
Vaccination complication	81.70570956	1.58E-19
Rash	68.93363732	1.02E-16
Anxiety	56.33309817	6.12E-14
Paraesthesia oral	51.40599746	7.51E-13
Dermatitis allergic	50.13558624	1.43E-12
Oxygen saturation	49.73241883	1.76E-12
Flushing	49.3121747	2.18E-12
Allergy to vaccine	44.76216274	2.22E-11
Heart rate increased	41.07021225	1.47E-10
Electrocardiogram normal	40.11780423	2.39E-10
Palpitations	37.25210863	1.04E-09
Dysphonia	36.7245365	1.36E-09
Erythema	34.31261596	4.69E-09
Oxygen saturation normal	33.65197646	6.59E-09
Cough	33.12717418	8.63E-09
Electrocardiogram	32.54342042	1.17E-08
Chest pain	31.8973366	1.63E-08
Eye pruritus	31.06355091	2.50E-08
Oedema peripheral	28.64424038	8.70E-08

MedDRA term	Chi-square	P-value
Heart rate	28.6141026	8.83E-08
Oral pruritus	28.13879224	1.13E-07
Idiopathic urticaria	26.77190018	2.29E-07
Angioedema	24.88169145	6.10E-07
Tachycardia	24.1470991	8.93E-07
Ocular hyperaemia	23.56285888	1.21E-06
Dizziness	21.90501031	2.86E-06
Pruritus generalised	20.41537534	6.23E-06
Hyperventilation	20.28914823	6.66E-06
X-ray normal	18.62066883	1.59E-05
Rash erythematous	17.79906026	2.46E-05
Chest X-ray normal	17.02743339	3.68E-05
Non-cardiac chest pain	16.87767466	3.99E-05
Oxygen saturation decreased	16.50541696	4.85E-05
Adverse drug reaction	15.84086837	6.89E-05
Asthma	14.94011526	0.000110978
Hypertension	13.76604066	0.000207045
Rhinitis	13.68760651	0.000215874
Food allergy	13.58133526	0.000228446
Rash macular	12.90979478	0.000326867
Blood glucose increased	12.39650931	0.000430137
Bronchial hyperreactivity	11.95078269	0.000546244
Oedema mouth	11.95078269	0.000546244
Dry throat	11.78175253	0.000598141
Respiratory rate	11.513761	0.000690829
Chest X-ray	10.74988156	0.00104286
Paraesthesia	10.46235549	0.001218318
Tension	9.777425891	0.001766675
Pyrexia	9.460175584	0.00209981
Feeling abnormal	9.424379867	0.002141195
Presyncope	9.414183846	0.002153134
Altered state of consciousness	9.010832195	0.002683842
Respiratory rate decreased	9.010832195	0.002683842
Rhinitis allergic	9.010832195	0.002683842
Red blood cell count normal	9.010832195	0.002683842
Respiration abnormal	9.010832195	0.002683842
Skin test	9.010832195	0.002683842
X-ray	8.958551384	0.002761738
Eyelid oedema	8.395515718	0.003761478
Hypoaesthesia oral	8.260899726	0.004050805
Feeling hot	8.222546332	0.004137311
Face oedema	8.081313552	0.004472402

MedDRA term	Chi-square	P-value
Immediate post-injection reaction	7.72106081	0.005458032
Blood glucose	7.72106081	0.005458032
Stridor	7.064359153	0.007863244
No reaction on previous exposure to drug	6.745371787	0.009399119
Blood pressure	5.971226848	0.014541161
Dermatitis	5.813875639	0.015900215
Feeling jittery	5.685593271	0.017104755
Lymph node palpable	5.624025895	0.017715909
Activated partial thromboplastin time shortened	5.624025895	0.017715909
Panic disorder	5.624025895	0.017715909
Skin test negative	5.624025895	0.017715909
Arrhythmia supraventricular	5.624025895	0.017715909
Steroid therapy	5.624025895	0.017715909
Oropharyngeal spasm	5.624025895	0.017715909
Soft tissue inflammation	5.624025895	0.017715909
Laryngospasm	5.624025895	0.017715909
Vaccination site erythema	5.624025895	0.017715909
Barium swallow normal	5.624025895	0.017715909
Lip discolouration	5.624025895	0.017715909
Plantar fasciitis	5.624025895	0.017715909
Food aversion	5.624025895	0.017715909
Computerised tomogram thorax normal	5.624025895	0.017715909
Oropharyngeal swelling	5.624025895	0.017715909
Vaccination site pruritus	5.624025895	0.017715909
Scan myocardial perfusion normal	5.624025895	0.017715909
Vasoconstriction	5.624025895	0.017715909
Blood electrolytes decreased	5.624025895	0.017715909
Venous thrombosis	5.624025895	0.017715909
Troponin	5.474670434	0.019293998
Pain in extremity	5.123595971	0.023602658
Bronchitis	4.782141775	0.028756334
Myalgia	4.763685188	0.029066252
Blood pressure decreased	4.744564425	0.029390993
Metabolic function test	4.672009897	0.030658028
Oxygen supplementation	4.300705957	0.038096556
Productive cough	4.18625334	0.040753069
Serum sickness	3.874258039	0.049031977
Hypokalaemia	3.874258039	0.049031977
Bronchospasm	3.874258039	0.049031977
Hypoventilation	3.874258039	0.049031977

Appendix H

Summary of the Seeker collaboration work and results

BRINKMAN REPORT

Brinkman-Seeker Collaboration 2012
Automated Classification of Adverse Events

seeker

Table of CONTENTS

ABSTRACT

The field of medical informatics has become an important area of research in the healthcare industry. This unique field unites researchers with backgrounds in computer science, engineering, the life sciences and healthcare. Due to this diverse set of skill requirements, now more than ever, strong partnerships between academia and industry are needed to develop efficient and intelligent solutions for a wide variety of healthcare issues.

In this pilot study, Seeker developers partnered with researchers from the BC Cancer Agency to investigate the task of identifying adverse events following immunizations using machine learning classification with simple language features. While previous work demonstrated that more advanced feature engineering is required for the identification of structurally complex adverse events, the results of this pilot study find that simple features perform well in some circumstances, and warrant further investigation and collaborative research.

More importantly, the partnership between Seeker and the BC Cancer Agency demonstrates a successful dialogue between industry partners and academic researchers, and shows how fruitful collaborative work can be in the medical informatics domain.

BACKGROUND

Public health authorities across North America are searching for ways to improve the safety and cost efficiency of many healthcare system components. One interesting and important area of public health focuses on the incidence of adverse events following an immunization (AEFI).

As defined by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, an adverse event (AE) is:

> *Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment. An adverse event (AE) can therefore be any unfavourable and unintended sign (including an abnormal laboratory finding, for example), symptom, or disease temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product.*[1]

In Canada, the Canadian Adverse Event Following Immunization Surveillance System (CAEFISS) exists to monitor the frequency and severity of AEFIs, and provides valuable data to help public health authorities make decisions related to immunization programs[2].

The process to submit an AEFI report to CAEFISS involves several steps, as indicated in Figure 1. When an AEFI occurs, a health care provider such as a nurse or physician compiles a report, and submits it to their local Provincial or Territorial Health Unit. The exact format and content of the AEFI report varies based on the standards and processes established by the Province or Territory, and does not necessarily mirror the fields and format of the nationally available AEFI report form[3] provided by the Public Health Agency of Canada (PHAC).

It is important to note that the reporting clinician provides immediate treatment of the adverse event prior to submitting the AEFI report to their local Health Unit, ensuring that the patient receives timely resolution of their symptoms. Once collected, Provincial and Territorial Health Units remove personally identifiable information from the report, and forward it to PHAC to be included in CAEFISS for aggregation and study.

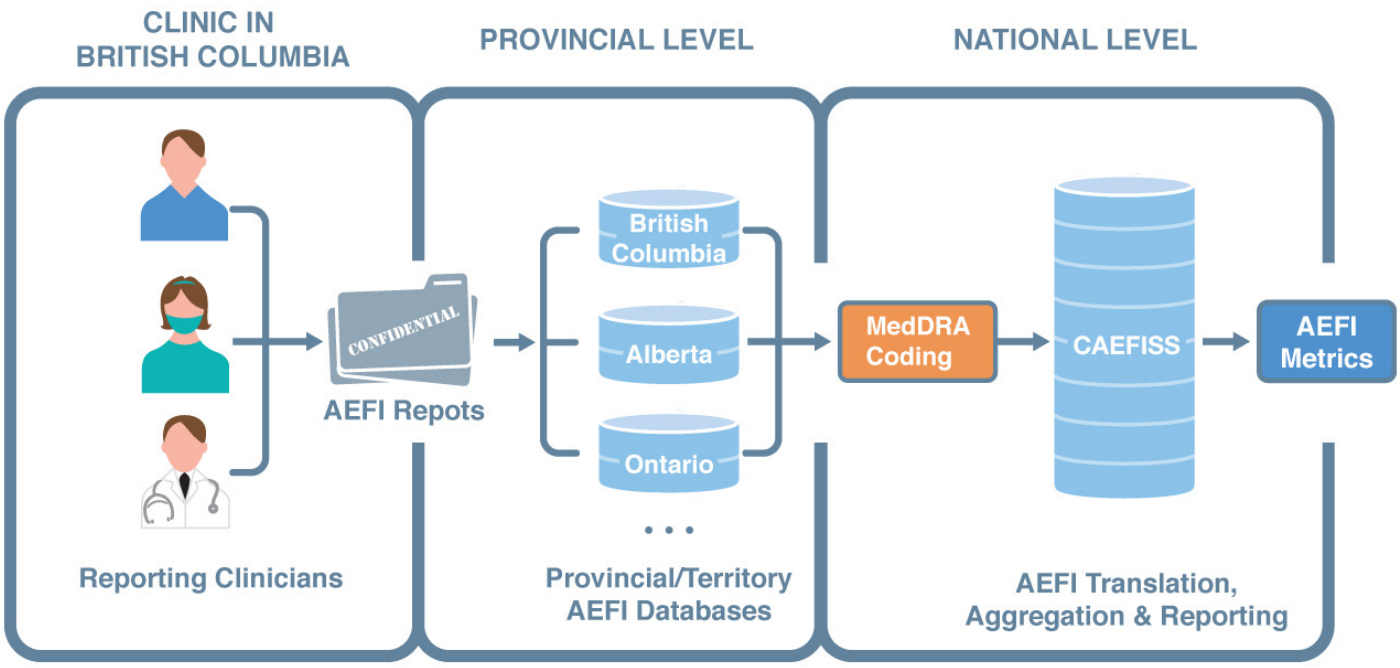


Figure 1. Information flow from Provincial / Territorial AEFI reporting systems to CAEFISS.

THE CHALLENGE

While AEFI report forms contain highly structured fields, there are sections that allow for the input of free text as supplementary information. This type of supplementary information is extremely valuable since its proper analysis could be used to improve the consistency and accuracy of the structured fields of the AEFI report. In turn, these improvements could directly impact the quality of the decisions public health authorities make related to immunization programs and protocols. Free text analysis of the supplementary information fields is where Seeker Solutions Inc. (Seeker) decided to explore the application of Natural Language Processing (NLP) and Machine Learning (ML) technologies.

In late 2012, a team of Seeker data scientists partnered with researchers from the BC Cancer Agency: Dr. Ryan Brinkman (Associate Professor, Medical Genetics, UBC; Senior Scientist, BC Cancer Agency) and Mélanie Courtot (PhD Candidate, UBC). Their goal was to determine if simple NLP and ML techniques and tools could be used to identify AEFIs within the free text fields of an AEFI report, and to tentatively identify a scale of difficulty for computationally identifying different types of AEFIs.

METHODOLOGY

Data Availability

A key issue for the project was to identify and obtain data that could be used for testing and proof of concept. Given the tight timeline to produce a proof of concept, Mélanie Courtot suggested that the team analyze data sets derived from the United States Vaccine Adverse Event Reporting System (VAERS). Similar to CAEFISS, VAERS is a national program designed to collect AEFI reports for the purposes of post-market vaccine safety monitoring [4].

However, a key difference between CAEFISS and VAERS is that VAERS data is made available to the public after reports have been suitably anonymized. In addition, while the VAERS AEFI reporting forms differ from those used in Canada, the bulk of the form is composed of a free text field used to capture details about the AEFI. This wealth of free text provided a good starting point for Seeker to apply NLP and ML technology, given its similarity to the supplementary information fields found on Canadian AEFI reports. The dataset already contains annotations for many different adverse events, as defined by the Medical Dictionary for Regulatory Activities (MedDRA)[5]. Finally, medical officers from the U.S. Food and Drug Administration (FDA) manually reviewed and positively coded 237 reports for anaphylaxis.

Adverse Events Selected for Identification

Two different adverse events were selected for the study:

1. **Paresthesia:** a burning or prickling sensation usually felt in the feet and hands that is idiomatically described as “tingling” or “pins and needles”[6].

2. **Anaphylaxis:** a severe, life-threatening, multi-system allergic reaction that occurs after contact with an allergen that may include some compounds found in a vaccine[7].

To diagnose anaphylaxis with various levels of certainty, the Brighton Collaboration Allergic Reactions Working Group has produced a case definition that describes symptoms that must be present in the dermatologic, cardiovascular, gastrointestinal, and respiratory systems[8]. For example, at the first level of diagnostic certainty, Brighton defined criteria must be present from a dermatological system, combined with symptoms present in a cardiovascular and/or respiratory system. Therefore, a physician may use terms such as “throat” and “swell” to describe the respiratory distress experienced by a patient, and “rash” and “hives” to describe their dermatological symptoms.

According to the Brighton case definitions, both groups of symptoms must be present and have appeared with a sudden onset and rapid progression before a diagnosis of anaphylaxis can be certain. To the best of our knowledge, no similar case definition exists for paresthesia.

METHODOLOGY

The Technology

The act of ***classifying*** a report as positive or negative for a condition is a well-known task in the ML community. For example, a ***classifier*** can learn to identify suspected cases of anaphylaxis or paresthesia through ***empirical evidence.*** To do so, the classifier is provided a *training dataset* containing a large number of reports that have already been positively or negatively labeled. The classifier then constructs a model by associating ***features*** that appear in the dataset with the positive or negative label.

In a typical classification task, there may be hundreds or thousands of features that the classifier observes. Each feature reflects an interesting aspect of the data, such as the length of the document or the frequency of a word. Many features such as these can be discovered within a free text report using various NLP techniques.

Once the learning process is complete, a trained classifier can use features within a ***test dataset*** (or in ***novel data***) to make positive or negative predictions against its constructed model. Past work has revealed that classifying anaphylaxis based on free text is challenging[9]. This is due to the variety of language that can be used to describe various systemic reactions, combined with a strict set of requirements for their valid configurations. Botsis *et al.* demonstrated that the production of a highly accurate classifier involves advanced ***feature engineering*** to incorporate medical domain knowledge into the classifier.

However, in many similar ML classification tasks, simple approaches have historically yielded decent results. Thus Seeker's approach to the classification problem was to use individual words as features (known as a ***bag-of-words***) instead of investing in advanced feature engineering. From an NLP standpoint, constructing a bag-of-words involves little domain knowledge, is computationally inexpensive, and requires little time to construct a classifier.

METHODOLOGY

The Experiments

To classify paresthesia, 32,885 reports were selected from the VAERS database between January 1, 2009 and December 31, 2009. Of these reports, 1,167 contained a MedDRA annotation for paresthesia. For each report, a bag-of-words was created. Each word was stemmed such that morphological inflections were removed (for example, the word ***infected*** would be stemmed to ***infect***). A ***10-fold cross validation*** was then performed on the dataset.

The cross validation technique works by randomly generating 10 individual views or ***folds*** of the data such that each fold reserves 90% of the data for training and 10% of the data for testing. For each fold, a classifier was trained using the training data for the fold, while performance metrics were collected using the testing data for the fold. Aggregate performance metrics were compiled from the results of each fold, and included the ***positive predictive value*** (known in the ML community as ***precision***), ***sensitivity*** (known in the ML community as ***recall***), and ***F-measure.***

In terms of classifiers, a variety of well-known types were used in the experiment, including Support Vector Machines, Random Forests, and Logistic Regression. A similar process was used to classify anaphylaxis, making use of 6,034 reports from the VAERS database. Of these reports, 237 were positively coded for anaphylaxis by the FDA. Some pre-processing on the FDA data was necessary to match the records back to their original VAERS report IDs.

RESULTS

For the paresthesia classification task, the calculated precision following a 10-fold cross validation ranged from 51 - 89% across the bundle of classifiers used. Recall ranged from 73 - 79%, and F-measure ranged from 62 - 80%. Overall, the most performant classifier model had a precision of 88%, a recall of 73%, and an F-measure of 80%. While preliminary in nature, these results demonstrate that simple NLP and ML techniques can be used to obtain relatively good results for AEFIs such as paresthesia, and that not all AEFIs require deep domain knowledge for identification.

In terms of a scale of difficulty, we consider these types of AEFIs to be relatively ***uncomplicated***, due mainly to the fact that inexpensive feature engineering (such as the bag-of-words approach) results in features that are sufficiently informative to produce a good classifier.

In practical terms, this preliminary finding is fairly significant, since it suggests that classifiers for uncomplicated AEFIs will remain relatively cheap to build and deploy. However, further work is needed for an in-depth error analysis, and to close the performance gap in precision and recall.

The anaphylaxis classification task tells a different story. As expected, the simple bag-of-words approach performed poorly when compared to the results obtained by Botsis *et al.* This is likely due to its inability to model medically significant domain knowledge.

In terms of performance, precision ranged from 29 - 61%, recall ranged from 18 - 61%, and F-measure ranged from 28 - 46%. Overall, the most performant classifier had a precision of 42%, a recall of 50%, and an F-measure of 46%.

On a scale of difficulty, we consider AEs similar to anaphylaxis to be structurally complex due to the fact that many simple features need to be combined or parsed according to a domain-specific recipe to generate more informative ones. More work similar to Botsis *et al.* is needed to understand what medical domain knowledge is needed to build highly accurate classifiers for AEFIs that are ***structurally complex***, as well as work to discover other informative features that may be useful for their identification and classification.

GOING FORWARD

The results of the experiments demonstrate that the difficulty of identifying various AEFIs can vary widely, and advanced feature engineering is not always required. However, much more work is needed to understand the structure of other AEFIs, and to find informative features that exist in the text to help with their classification and discovery. Further work is also needed to close the performance gap in precision and recall for both uncomplicated and structurally complex AEFIs.

Overall, given the wide range of AEFIs, collaboration between the ML and NLP communities with domain experts will continue to be a necessity. This work demonstrates how fruitful these types of collaborations can be, from the discovery of new data sources to the transfer of domain knowledge between academia and industry. While the duration of the partnership was short, both parties agree that the simple experiments and their preliminary results represent promising potentials for complex algorithmic processing in the medical domain.

REFERENCES

1. *Clinical Safety Data Management: Definitions and Standards for Expedited Reporting E2A.* ICH Harmonised Tripartite Guideline. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E2A/Step4/E2A_Guideline.pdf (Retrieved November 8, 2013)
2. *Canadian Adverse Events Following Immunization Surveillance System (CAEFISS).* Public Health Agency of Canada. http://www.phac-aspc.gc.ca/im/vs-sv/caefiss-eng.php (Retrieved October 17, 2013)
3. *Report of Adverse Events Following Immunization (AEFI).* Public Health Agency of Canada. http://www.phac-aspc.gc.ca/im/pdf/raefi-dmcisi-eng.pdf (Retrieved October 10, 2013)
4. *VAERS: Vaccine Adverse Event Reporting System.* http://vaers.hhs.gov/index (Retrieved October 17, 2013)
5. *Understanding MedDRA - The Medical Dictionary for Regulatory Activities.* MedDRA. http://www.meddra.org/sites/default/files/page/documents/meddra2013.pdf (Retrieved November 8, 2013)
6. *NINDS Paresthesia Information Page.* National Institute of Neurological Disorders and Stroke. http://www.ninds.nih.gov/disorders/paresthesia/paresthesia.htm (Retrieved October 17, 2013)
7. M. Erlewyn-Lajeunesse, et al. (2007) *Anaphylaxis as an adverse event following immunisation.* Journal of Clinical Pathology, 60(7), pages 737-739. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1995783/ (Retrieved October 30, 2013)
8. J.U. Ruggeberg et al. (2007) *Anaphylaxis: Case Definition and Guidelines for Data Collection, Analysis, and Presentation of Immunization Safety Data.* Vaccine, 25(31), pages 5675-5684. http://dx.doi.org/10.1016/j.vaccine.2007.02.064
9. T. Botsis *et al.* (2011) *Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection.* Journal of the American Medical Association, 18, pages 631-638. http://dx.doi.org/10.1136/amiajnl-2010-000022

Seeker Solutions

Corporate Head Office
400-1112 Fort St.
Victoria, BC V8V 3K8
250.483.4129

seeker

www.seekersolutions.com