Ontology-based methods for analyzing life science data

Olivier Dameron

UNIVERSITÉ DE RENNES 1

Habilitation à Diriger des Recherches

présentée par

Olivier Dameron

Ontology-based methods for analyzing life science data

Soutenue publiquement le 11 janvier 2016

devant le jury composé de

Anita Burgun	Professeur, Université René Descartes Paris	Examinatrice
Marie-Dominique Devignes	Chargée de recherches CNRS, LORIA Nancy	Examinatrice
Michel Dumontier	Associate professor, Stanford University USA	Rapporteur
Christine Froidevaux	Professeur, Université Paris Sud	Rapporteure
Fabien Gandon	Directeur de recherches, Inria Sophia-Antipolis	Rapporteur
Anne Siegel	Directrice de recherches CNRS, IRISA Rennes	Examinatrice
Alexandre Termier	Professeur, Université de Rennes 1	Examinateur

Contents

Acknowledgments

I and indebted to Mark Musen, Anita Burgun and Anne Siegel for allowing me to join their team, for their help and their encouragement, as well as for setting up such exciting environments.

I am most grateful to Michel Dumontier, Christine Froidevaux and Fabien Gandon for kindly accepting to review this document, and also to Anita Burgun, Marie-Dominique Devignes, Anne Siegel and Alexandre Termier for accepting to be parts of the committee.

I am thankful to all my colleagues for all the work we have accomplished together, as summarized by Figure 1 on the facing page. The PhD students I co-supervised have brought significant contributions to most of the works presented in this manuscript: Nicolas Lebreton, Élodie Roques, Charles Bettembourg, Philippe Finet, Jean Coquet and Yann Rivault.

Close collaborations with colleagues from whom I learnt a lot were very stimulating: Daniel Rubin, Natasha Noy, Anita Burgun, Julie Chabalier, Andrea Splendiani, Paolo Besana, Lynda Temal, Pierre Zweigenbaum, Cyril Grouin, Annabel Bourdé, Oussama Zekri, Pascal van Hille, Jean-François Éthier, Bernard Gibaud, Régine Le Bouquin-Jeannès, Anne Siegel, Jacques Nicolas, Guillaume Collet, Sylvain Prigent, Geoffroy Andrieux, Nathalie Théret, Fabrice Legeai, Anthony Bretaudeau, Olivier Filangi and again Charles Bettembourg.

Collaborating with colleagues from INRA was a great opportunity to tackle "real problems" and gave a great perspective on what this is all about: Christian Diot, Frédéric Hérault, Denis Tagu, Mélanie Jubault, Aurélie Évrard, Cyril Falentin Pierre-Yves Lebail and all the ATOL curators: Jérôme Bugeon, Alice Fatet, Isabelle Hue, Catherine Hurtaud, Matthieu Reichstadt, Marie-Christine Salaün, Jean Vernet, Léa Joret. Similarly, collaborations with colleagues from EHESP provided the "human" counterpart: Nolwenn Le Meur, Yann Rivault.

More broadly, I benefited from the interactions with many other colleagues: Gwenaëlle Marquet, Fleur Mougin, Ammar Mechouche, Mikaël Roussel, as well as the whole Symbiose group at IRISA, and particularly Pierre Peterlongo.

Working on workflows with Olivier Collin, Yvan Le Bras, Alban Gaignard and Audrey Bihouée has been fun and I am eager to see what will come out of it.

In addition to research all these years were also the occasion of great teaching-related encounters: Christian Delamarche, Emmanuelle Becker, Emmanuel Giudice, Annabelle Monnier, Yann le Cunff, Cédric Wolf.

Thank you all, I enjoyed all of it.

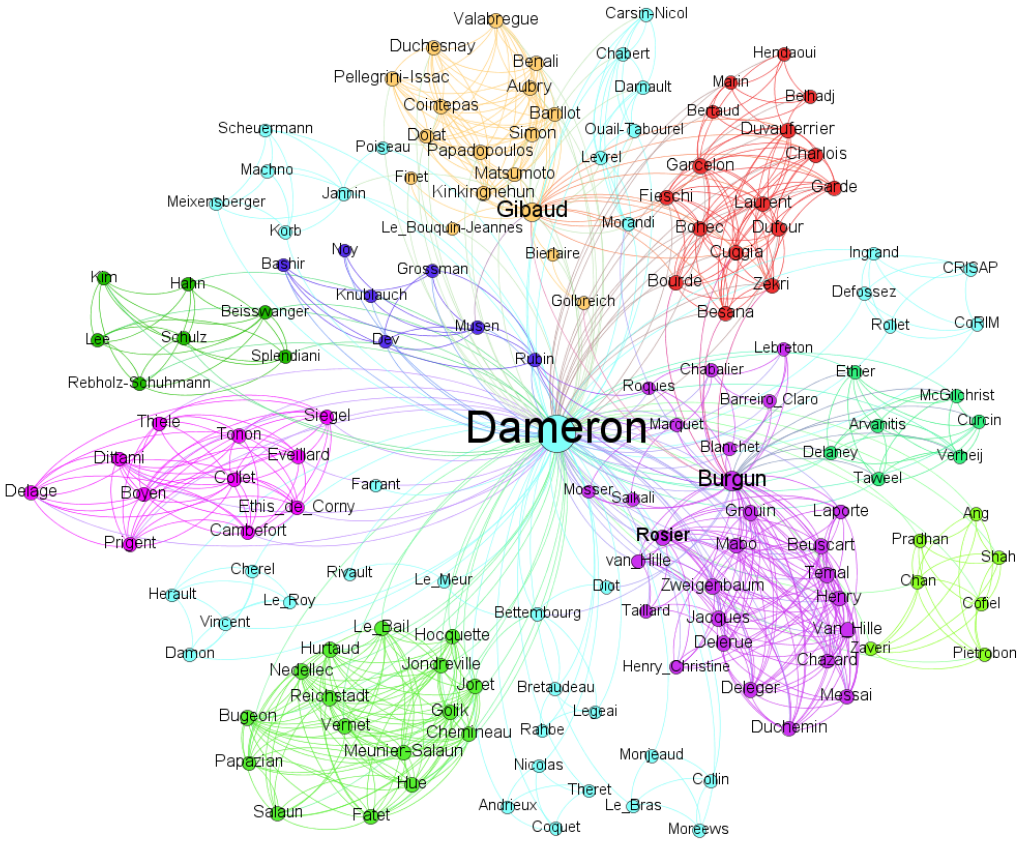

Figure 1: Graph of my co-authors. Two authors are linked if they share at least one publication. Node size is proportional to the number of articles.

Chapter 1

Introduction

This document summarizes my research activities since the defense of my PhD in December 2003. This work has been carried initially as a postdoctoral fellow at Stanford University with Mark Musen's Stanford Medical Informatics group (now BMIR[1]), and then as an associate professor at University of Rennes 1, first with the UPRES-EA 3888 (which became UMR 936 INSERM – University of Rennes 1 in 2009) from 2005 to 2012, and then with the Dyliss team at IRISA since 2013.

First, I will present the context in which my research takes place. We will see that the traditional approaches for analyzing life science data do not scale up and cannot handle their increasing quantity, complexity and connectivity. It has become necessary to develop automatic tools not only for performing the analyses, but also for helping the experts do it. Yet, processing the raw data is so difficult to automate that these tools usually hinge on annotations and metadata as machine-processable proxies that describe the data and the relations between them.

Second, I will identify the main challenges. While generating these metadata is a challenge of its own that I will not tackle here, it is only the first step. Even if metadata tend to be more compact than the original data, each piece of data is typically associated with many metadata, so the problem of data quantity remains. These metadata have to be reused and combined, even if they have been generated by different people, in different places, in different contexts, so we also have a problem of integration. Eventually, the analyses require some reasoning on these metadata. Most of these analyses were not possible before the data deluge, so we are inventing and improving them now. This also means that **we have to design new reasoning methods for answering life science questions using the opportunities created by the data deluge while not drowning in it**. Arguably, biology has become an information science.

Third, I will summarize the contributions presented in the document. Some of the reasoning methods that we develop rely on life science background knowledge. Ontologies are the formal representations of the symbolic part of this knowledge. The Semantic Web is a more general effort that provides an unified framework of technologies and associated tools for representing, sharing, combining metadata and pairing them with ontologies. **I developed knowledge-based reasoning methods for life science data.**

Finally, I will describe the organization of the manuscript.

[1] `http://bmir.stanford.edu/`

1.1 Context: integrative analysis of life science data

Life sciences are intrinsically complicated and complex [1, 2]. Until a few years ago, both the scarcity of available information and the limited processing power imposed the double constraints that work had to be performed on fragmented areas (either precise but narrow or broad but shallow) as well as using simplifying hypotheses [3].

The recent joint evolution of data acquisition capabilities in the biomedical field, and of the methods and infrastructures supporting data analysis (grids, the Internet...) resulted in an explosion of data production in complementary domains (*omics, phenotypes and traits, pathologies, micro and macro environment...) [3, 4, 5]. For example, the BioMart community portal provides a unified interface to more than 800 biological datasets distributed worldwide and spanning genomics, proteomics and cancer data [6], and the 2015 *Nucleic Acids Research* Database issue refers to more than 1500 biological databases [7]. Making data reusable has been widely advocated [8]. This "data deluge" is the life-science version of the more general "big data" phenomenon, with the specificities that the proportion of generated data is much higher, and that these data are highly connected [9].

In addition to the breakthrough in each of these domains, majors efforts have been undertaken for developing the links between them: systems biology[2] [10, 11, 12, 13, 14] at the fundamental level, translational medicine[3] [15, 16] for the link between the fundamental and clinical levels, and more recently translational bioinformatics[4] [17, 18, 19, 20] for the link between what happens at the molecular and cellular levels and what happens at the organ and individual levels. These links between domains are obviously useful for performing better analyses of data, but conversely these new connections can sometimes reshape the domains themselves. For example, translational bioinformatics modifies the definitions of the fundamental notion of what constitutes a disease by considering sequencing of genes or quantitating panels of RNA in addition to the traditional nosology [21].

We are witnessing the transition from a world of isolated islands of expertise to a network of inter-related domains [4, 22]. This is supported by another transition from a world where we had a small quantity of informations on a lot of people to a world where we have a lot of informations in related domains (genetics, pathology, physiology, environment) for a small but increasing number of people. Storage capabilities kept pace with the increasing data generation. However, **the bottleneck that once was data scarcity now lies in the lack of adequate data processing and data analysis methods**. This increasing data quantity and connectivity was the origin of new challenges.

The stake of data integration consists in establishing and then using systematically the links between elements from different domains (e.g. from *omics to pathologies, from pathologies to *omics, or between *omics or pathologies of different species) having potentially different levels of precision [5]. For example, meta-analysis of heterogeneous annotations and pre-existing knowledge often lead to novel insights undetectable by individual analyses [23, 24]. Systems

[2]**Systems biology** aims at modeling the interactions between the elements of a biological system and their emergent properties. These elements can themselves be composed of sub-elements that can interact among them or with other elements.

[3]**Translational medicine** aims at providing the best treatment for each patient by using the most recent discoveries in biology, drug discovery and epidemiology (bench to bedside), and conversely to reuse medical data when performing research (bedside to bench).

[4]**Translational bioinformatics** derives from translational medicine and focuses on integrating information on clinical and molecular entities. It aims at improving the analysis and affect clinical care.

biology, translational medicine and translational bioinformatics all focus on the systematic organization of these links.

The systematic exploitation of data permitted by integration requires some kind of automation. Because of life sciences intrinsic complexity, vast quantities of elements as well as the numerous links between them that represent their inter-dependencies have to be taken into account [25, 26].

This systematic exploitation of data is not only massive, it is also complex [4, 27]. **The systematic analysis of the integrated data requires to perform some interpretation, which hinges on background knowledge** [3]. Expertise or domain knowledge can be seen as the set of rules representing in what conditions data can be used or can be combined for inferring new data or new links between data (Levesque also provided an excellent more general article on knowledge representation for artificial intelligence [28]).

The remainder of this document focuses on the third challenge of using knowledge for automatically integrating and analyzing biomedical data in a context covering translational medicine and translational bioinformatics.

1.2 Challenges: using domain knowledge to integrate and analyze life science data

Several bottlenecks hamper the automated systematic exploitation of biomedical data:

- **it has to take expertise or knowledge into account** [29]. This entails both to represent such knowledge in a formalism supporting its use in an automatic setting, and that the conditions determining knowledge validity are themselves formally represented.
- **it relies on data and knowledge that are obviously incomplete** [3, 30]. We are therefore in the intermediate state where we must develop automatic methods for processing vast amount of heterogeneous and inter-dependent data while being limited by the incomplete and fragmentary aspect of these data.
- **it produces results that are so big and so complex that their biological interpretation is at best difficult**. Dentler et al. showed that "Today, clinical data is routinely recorded in vast amounts, but its reuse can be challenging" [31]. Moreover, it is not only the quantity of data that is increasing, but also the associated metadata that describe and connect these data. Rho et al. point out that "One important issue in the field is the growing complexity of annotation data themselves" and that "Major difficulties towards meaningful biological interpretation are integrating diverse types of annotations and at the same time, handling the complexities for efficient exploration of annotation relationships" [24].

As Stevens et al. noted, "Much of biology works by applying prior knowledge [...] to an unknown entity, rather than the application of a set of axioms that will elicit knowledge. In addition, the complex biological data stored in bioinformatics databases often require the addition of knowledge to specify and constrain the values held in that database" [29]. The same holds for the biomedical domain, e.g. to identify patient subgroups in clinical data repositories [32].

The knowledge we are focusing on is mostly symbolic, as opposed to other kinds of biomedical knowledge (probabilistic, related to chemical kinetics, 3D models of anatomical entities or 4D models of processes...). It should typically support generalization, association and deduction.

There is a long tradition of works in order to come up with an explicit and formal representation of this knowledge that would support automatic processing. Cimino identified the following key requirements: "vocabulary content, concept orientation, concept permanence, non-semantic concept identifiers, polyhierarchy, formal definitions, rejection of 'not elsewhere classified' terms, multiple granularities, multiple consistent views, context representation, graceful evolution, and recognized redundancy" [33, 34, 35].

This line of work resulted in the now widespread acceptance of ontologies [29, 36] to represent the biomedical entities, their properties and the relations between these entities. Bard et al. defined **ontologies** as "formal representations of knowledge in which the essential terms are combined with structuring rules that describe the relationships between the terms" [37]. This covers the main points and encompasses alternative definitions [38, 39].

Ontologies range from fairly simple hierarchies to semantically-rich organization supporting complex reasoning [36]. There is also a distinction depending on their scope. Top-level ontologies (or upper ontologies) such as DOLCE or BFO are domain-independent and represent general notions such as things and processes. Domain ontologies cover a specific domain (e.g. normal human anatomy for the FMA of the description of gene products for GO) [40].

Ontologies are now a well established field [36, 2] that evolved from concept representation [41]. In May 2015, there were 442 ontologies referenced by BioPortal, and 10,768 PubMed articles mentioning "ontology" (Figure 1.1 on the next page). They cover the creation of new ontologies, data annotation [2], data integration [3, 42], data analysis [5], or ontology as a proper research field [43]. There are many applications for bio-ontologies themselves, for example analysis of cancer *omics data [44], integration and analysis of quantitative biology data for hypothesis generation [45], biobanks [42], interpretation of complex biological networks [46] or even analysis of research funding by diseases [47]. Hoehndorf et al. recently performed a review of the importance of bio-ontologies and their main application domains [48]. Among the main ontologies are the Gene Ontology (GO; for an analysis of its becoming the most cited ontology, see [45]), that provides a species-independent vocabulary for describing gene products, the NCI thesaurus for describing cancer-related entities, the International Classification of Diseases (ICD), OMIM for human genetic disorders, SNOMED Clinical Terms, the US National Drug File (NDF-RT), ChEBI for describing small chemical molecules and UNIPROT for describing proteins, the Medical Subject Headings (MeSH) for annotating PubMed articles [36].

As noted previously, there are now numerous ontologies that are used in various contexts. These ontologies can overlap, which hampers data integration as some resources refer to some entity in an ontology whereas other resources can refer to the corresponding entity in another ontology. The Unified Medical Language System (UMLS) provides some unifying architecture between the major biomedical ontologies and terminologies. The problems of ontology dispersion and overlap found a solution with BioPortal[5] [49]. It is an open repository of biomedical ontologies that offers the possibility to browse, search and visualize ontologies, as well as to create, to store and to use mappings between these ontologies (i.e. relations between entities from different ontologies). Bioportal also supports the annotation of data from Gene Omnibus, clinical trials and ArrayExpress. It should be noted that BioPortal also provides some API and Web services, so it can also be used by programs.

[5] `http://bioportal.bioontology.org/`

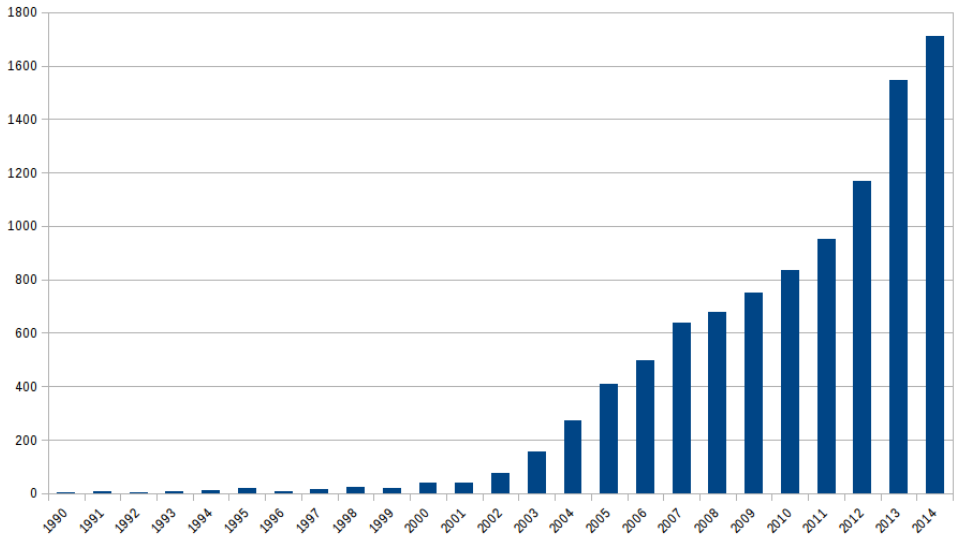

Figure 1.1: Evolution of the number of PubMed articles referencing "ontology".

The emergence of ontologies in biomedical informatics and bioinformatics happened in parallel with the development of the **Semantic Web** in the computer science community [50, 41]. The Semantic Web is an extension of the current Web that recognizes the need to represent data on the Web in machine-processable formats and to combine them with ontologies. It aims to support fine-grained data representation for automatic retrieval, integration and interpretation. To do so,

- it shifts the granularity from document to each atomic data they contain, identifying them with specific URIs (now IRIs);
- it represents explicitly the relations linking some of these data by also designating them with their URIs (whereas we only have untyped `href` between HTML pages);
- it also encompasses the representation of generalities and of relations between them (e.g. Alzheimer's disease is a kind of neurodegenerative disease), as well as the connection between atomic data (which are anecdotal) and generalities (which are universal) so that a patient's disease with all its specificities can be described as an element of the set of the Alzheimer's diseases.

The W3C defines several recommendations (that are *de facto* standards) related to the Semantic Web initiative for data representation, integration and analysis. RDF[6] (Resource Description Framework) represents data and their relations using triples of URIs (the first designates the subject that is described, the second represents the relation or predicate, and the third represents the value of the relation for the subject, and is called the object). RDF provides a special property (`rdf:type`) to represent the fact that some data identified by its URI is an instance of a general class. RDFS[7] (RDF schema) and OWL[8] (Ontology Web Language) provide sets of RDF entities with special and formal semantics to represent generalities (so ontologies). Therefore, RDFS and OWL statements are also RDF triples. RDFS allows to represent taxonomies, and OWL provides several profiles with a more formal semantics that

[6] `http://www.w3.org/RDF/`
[7] `http://www.w3.org/TR/rdf-schema/`
[8] `http://www.w3.org/2001/sw/wiki/OWL`

support richer reasoning. RDFS is well adapted to simple reasoning over large data sets, whereas OWL is adapted to more complex reasoning, at the cost of potentially longer computation times. These reasoning tasks are supported by several other recommendations. SPARQL[9] (SPARQL Protocol and RDF Query Language) is an SQL-like query language for RDF. Note that SPARQL1.1 can take most of RDFS semantics into account. OWL does not have a query language but it does not really need one either because OWL inferences consists mostly in determining whether a piece of data is an instance of a class, or whether a class is a subclass of another one. Additionally, SWRL[10] (Semantic Web Rule Language) allows to represent inference rules with variables. It should be noted that even if most bio-ontologies are represented in OWL, very few take advantage of the language expressivity. Most are RDFS ontologies disguised in OWL (which is possible because OWL is built on top of RDFS, e.g. all OWL classes are RDFS classes), even if it has been demonstrated that they would benefit from using OWL's additional semantics [51, 52, 53]

Life sciences are a great application domain for the Semantic Web [54, 55, 56] and several major teams are involved in both, particularly at the W3C Semantic Web Health Care and Life Sciences interest group (HCLSIG)[11]. Since 2008, the Semantic Web Applications and Tools for Life Sciences (SWAT4LS) workshop[12] (co-organized by Andrea Splendiani, who was a postdoc at U936) is an active event, along with conferences such as DILS, ISWC and ESWC. Semantic Web technologies have become an integral part of translational medicine and translational bioinformatics [5, 14]. Several works have showed how these technologies can be used to integrate genotype and phenotype informations and perform queries [57, 58]. More recently, Holford et al. proposed a Semantic Web framework to integrate cancer omics data and biological knowledge [44]. The Linked Data initiative [59] and particularly the Linked Open Data project promotes the integration of data sources in machine-processable formats compatible with the Semantic Web. Figure 1.2 on the facing page shows the importance of life sciences. In the past few years, this proved instrumental for addressing the problem of data integration [53, 60]. In this context, the Bio2RDF project[13] promotes simple conventions to integrate biological data from various origins [61, 62, 63]. Moreover, Semantic Web technologies support federated queries that gather and combine informations from several data sources [62]. The reconciliation of identifiers is further facilitated by initiatives such as identifiers.org [64].

1.3 Summary of the contributions

My contributions focused on methods for automatic analysis of biomedical data, based on ontologies and Semantic Web technologies. This section is organized chronologically for presenting how the various themes evolved and were applied to different projects. The remainder of the document is organized thematically.

My PhD dissertation consisted in the creation of an ontology of brain cortex anatomy [65, 66]. At this point, the added value of ontologies for data integration and for reasoning had been demonstrated by several major projects for many years. However, it was clear that developing ontologies was a difficult endeavor with a part of craftsmanship. Particularly, one had to keep track of multiple dependencies between classes [67, 68, 69]. There was also the perception that the automatic reasoning based on ontologies all too often had to be completed by *ad-hoc* pro-

[9] `http://www.w3.org/TR/sparql11-overview/`
[10] `http://www.w3.org/Submission/SWRL/`
[11] `http://www.w3.org/wiki/HCLSIG`
[12] `http://www.swat4ls.org/`
[13] `https://github.com/bio2rdf`

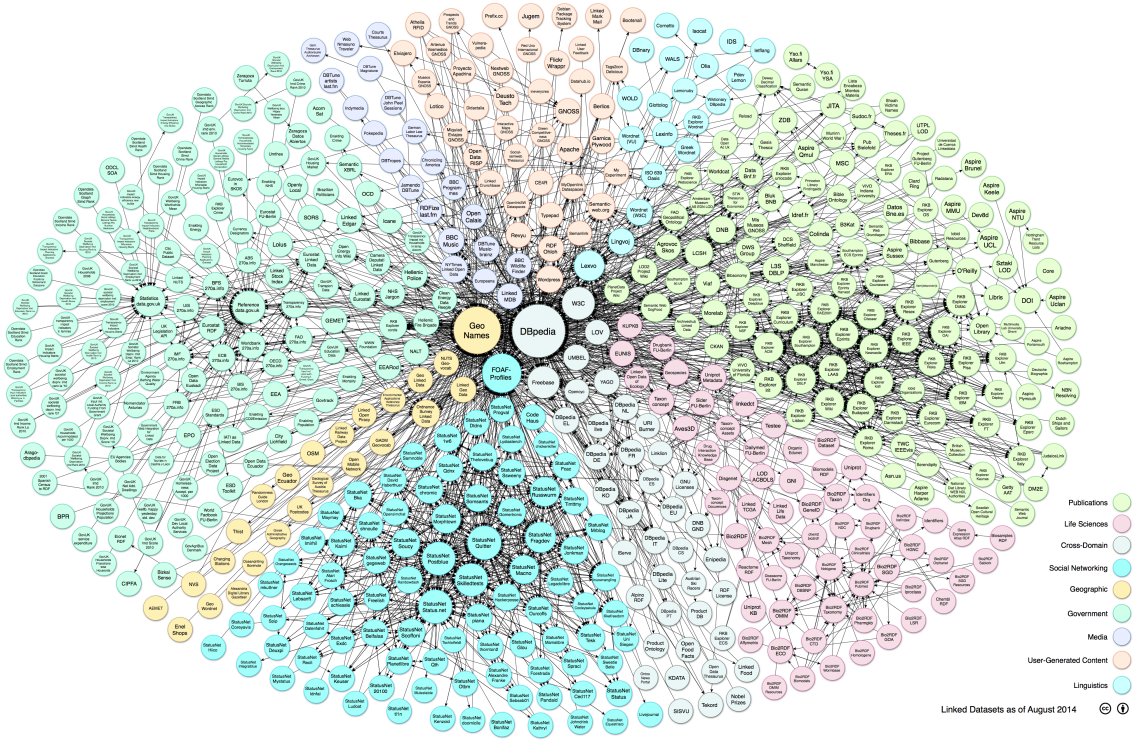

Figure 1.2: Linked Open Data cloud in August 2014. Nodes are resources. Edges are cross-references between resources. Life science resources constitute the purple portion in the lower right corner.
(`http://lod-cloud.net/`).

gramming extensions either as pre-processing for making the data amenable to reasoning, or as post-processing. There was no widespread agreement on the format to use for representing ontologies at the time: frames were the dominant paradigm but multiple implementations existed in addition to the Protégé editor [70]; interesting solutions like XML were advocated by the World Wide Web Consortium (W3C); Description Logics (DL) were gaining acceptance in the biomedical community thanks to the reasoning capability and the DAML+OIL format being associated with an open source editor and a reasoner [71].

By the time I started my post-doc at Mark Musen's Stanford Medical Informatics lab in Stanford University, DAML+OIL had evolved into the OWL effort, which became a formal W3C recommendation on February 2004. Holger Knublauch, also post-doc at SMI, had just started developing an OWL plugin for Protégé [72]. During my stay, I participated in the Virtual Soldier project.

My main contribution was to develop the symbolic reasoning mechanism for inferring the consequences of penetrating bullet injuries based on both anatomical knowledge and patient-specific data [73, 74]. This made extensive use of Description Logics expressiveness to leverage reasoning capabilities based on classes (generic reasoning for inferring that a class is a subclass of another one) as well as on instances (data-specific reasoning for inferring that an individual is an instance of a class). The reasoning relied on rich anatomical knowledge. The Foundational Model of Anatomy (FMA) was the reference ontology but was originally developed and maintained in frames, fortunately using Protégé.

My second contribution was on ontology modeling and representation. I studied the theo-

retical aspects of the conversion of the frame-based FMA into an OWL version by preserving as much as possible of its original semantic richness and by automatically adding features that were beyond frames expressivity such as necessary and sufficient definitions or disjointness constraints [75, 76]. Recognizing that different applications using the FMA may have different expressivity requirements and that some features may be useful in some context, but may add an unnecessary computational burden in other contexts, I proposed a modular approach so that users could import only the features they needed.

My third contribution addressed the need to automate certain operations during ontology development and ontology usage. For assisting both the development of the reasoning capability and the conversion of the FMA into OWL, I developed the ProtegeScript plugin [77] (still included in the distribution of Protégé 3) that added scripting capabilities (mainly Python, Ruby and BeanShell) to Protégé and was compatible with both the original frames setting and the OWL plugin. Eventually, I helped organize and teach the first versions of the Protégé Short Course and Protégé OWL Short Course in 2005, and have been invited back to Stanford to do so until 2011.

Since I joined Anita Burgun's team as an associate professor at Rennes 1 university, I continued working on ontology-based reasoning. Together with Gwenaëlle Marquet, we developed a semantically-rich reasoning application performing automatic classification of glioma tumors [78, 79]. This was in direct continuation of the line of work initiated in the Virtual Soldier project. In both cases, we demonstrated that if the relevant domain ontologies are rich enough, developing an application-specific reasoning module only requires the creation of a few classes. In both cases though, this assumption was optimistic. Many works by other teams focused on improving existing ontologies[2] such as GO [52] or the NCI Thesaurus [80, 81].

As Jim Hendler pointed out, even a little semantics goes a long way [82], and I extended my work to simpler forms of reasoning. With Julie Chabalier, we created a knowledge source relating diseases and pathways by integrating the Gene Ontology, KEGG orthology and SNOMED CT [83, 84, 85, 86]. We proposed an approach combining mapping and alignment techniques. We used OWL-DL as the common representation formalism and demonstrated that RDFS queries were expressive enough with acceptable computational performances. From 2008 to 2010, I supervised Nicolas Lebreton's PhD thesis with Anita Burgun on Web services parameters compatibility for semi-automatic Web Service composition [87, 88, 89]. The context was that biologists typically conduct the analysis of their results by building workflows of atomic programs that run on bioinformatics platforms and grids. They devote a great deal of efforts to building and maintaining (*ad-hoc*) scripts that execute these workflow and ensure the necessary data format conversions. We showed that the WSDL descriptions of Web services only provide a view on the structure of the services' input and output parameters, whereas a view on their nature was necessary for Web services composition. We proposed an algorithm using classes taxonomic hierarchies of Web services OWL-S semantic descriptions for checking the semantic compatibility of services parameters and for suggesting compatible parameters pairings between two Web services semi-automatically. We generated Taverna Xscufl files whenever possible. Lately, Charles Bettembourg developed GO2PUB as a part of his PhD thesis [90]. GO2PUB is a tool that uses the knowledge from the Gene Ontology (GO) and its annotations for enriching PubMed queries with gene names, symbols and synonyms. We used the GO classes hierarchy for retrieving the genes annotated by a GO term of interest, or one of its more precise descendants. We demonstrated that the use of genes annotated by either GO terms of interest or a descendant of these GO terms yields some relevant articles ignored by other tools. The comparison of GO2PUB, based on semantic expansion, with GoPubMed, based on text-mining techniques, showed that both tools are complementary.

With my participation to the Akenaton project, work on semantically-rich reasoning resumed and shifted to the optimization of symbolic knowledge modeling [91, 92]. The context is the automatic triage of atrial fibrillation alerts generated by implantable cardioverter defibrillators according to their severity. There can be up to twenty alerts per patient per day, with around 500,000 current patients en Europe, and an estimation of 10,000 new patients every year in Europe. Alerts severity depend on the CHA2DS2VASc score, which evaluation requires domain knowledge for reasoning about the patient's clinical context. Several modeling strategies are possible for reasoning on this knowledge. A first work compared ten strategies emphasizing all the possible combinations of Java, OWL-DL and SWRL to compute the CHA2DS2VASc score. A second work compared the best of these ten strategies with a Drools rules implementation [93]. The results showed that their limitations are the number and complexity of Drools rules and the performances of ontology-based reasoning, which suggested using the ontology for automatically generating a part of the Drools rules.

Together with the previous work on glioma tumor classification, the Akenaton project opened a new perspective on symbolic reasoning with incomplete information. When we were designing the reasoning module for grading glioma tumors, we observed that some information were missing for several patients, and that the module (rightfully) prevented the system from reaching a conclusion. We showed however that it was possible to narrow the number of possibilities by excluding the situations that we inferred to be impossible. In the Akenaton project, similarly, the CHA2DS2VASc score is computed by summing points for each criterion met by the patient. Missing information can result in under-estimating the real value of the CHA2DS2VASc score. We proposed as a complementary approach to start from the maximum possible CHA2DS2VASc score value and to subtract points for each criterion not met by the patient This in turn resulted in a over-estimation of the score. For a patient, combining the two values allowed us to determine the range of the possible scores, instead of the false sense of security provided by a value that may be under-estimated. Building on the experience, my contribution to the Astec project in 2012 was an OWL design pattern for modeling eligibility criteria that leveraged the open world assumption to address the missing information problem of prostate cancer clinical trial patient recruitment [94, 95].

Over the years, my interest in bioinformatics grew. In parallel with the previous works, I started in 2010 a collaboration with Christian Diot at UMR1348 PEGASE (INRA and Agrocampus Ouest) on knowledge-based cross-species metabolic pathway comparison in order to study how lipid metabolism was different in chicken human and mouse [96, 97]. Together, we supervised Charles Bettembourg's master in 2010 degree and ongoing PhD thesis since 2011. Our collaboration originated from the observation that when overfed, chicken do not develop liver steatosis, whereas other animals such as geese, mice and humans do. Liver steatosis can further evolve into fatty liver disease and cancers, so analyzing the specificities of chicken's lipid metabolism is of both agricultural and medical interest. Our approach is based on metabolic pathways structural comparison in order to identify common and species-specific reactions, and more importantly on functional comparison in order to quantify how much a metabolic process is common and species-specific. We improved a semantic similarity measure based on Gene Ontology and created another metric measuring semantic specificity. This work opened the opportunity of another collaboration with Frédéric Hérault on functional analysis and comparison of gene sets, where we demonstrated the benefits of using semantic similarity for post-processing and clustering DAVID results [98].

In 2013, I joined the Dyliss team at IRISA. I contributed to the analysis of the candidate metabolic networks for *Ectocarpus siliculosus* generated by Sylvain Prigent in the Idealg project during his PhD [99]. I am also working with Nathalie Théret, Geoffroy Andrieux and Jean

Coquet (whom I co-supervise) on the analysis of TGF-β signaling pathways and their role in human cancer [100]. Additionally, I collaborate with Fabrice Legeai, Anthony Bretaudeau, Charles Bettembourg and Denis Tagu, as well as with Mélanie Jubault and Aurélie Évrard on representing, storing and querying aphids [101] and *Brassicaceae* data in RDF. I co-supervise with Régine Le Bouquin-Jeannès and Bernard Gibaud from LTSI the PhD thesis of Philippe Finet on the integration and analysis of telemedicine data for monitoring patients with multiple chronic diseases [102, 103, 104]. I will co-supervise with Nolwenn Le Meur from EHESP the PhD thesis of Yann Rivault on the analysis of patients' care trajectories [105]. These works are still in progress and are further developed in my research perspectives in Chapter 6.

Over the years, the biomedical data and ontologies I have been using evolved from a medical/clinical context to more general biological one. However, the reasoning primitives remained the same, so the distinction is not really relevant. From this point, I will refer to life science data in general.

1.4 Organization of the manuscript

My various contributions belonged to different zones in the reasoning continuum ranging from the simple exploitation of a taxonomy to sophisticated reasoning involving intricate necessary and sufficient definitions and the open world assumption.

Chapter 2 presents reasoning based on hierarchy, which is valuable in spite of its simplicity as a way to circumvent computational limitations and because the task at hand does not require more elaborate features. Section 2.1 is a summary of RDF and RDFS principles and of the associated entailments. Section 2.2 emphasizes constraints due to ontologies' size and presents an early case study for inferring candidate associations between biological pathways from KEGG and diseases from SNOMED-CT using the Gene Ontology as a pivot. Section 2.3 focuses on a method for performing semi-automatic pairing of Web services parameters. Section 2.4 shows how computation performances support performing on the fly semantic expansion of PubMed queries.

Chapter 3 presents reasoning based on classification for inferring whether an individual is an instance of a class or whether a class is a subclass of another one. Section 3.1 is a summary of OWL main principles and the associated inferences. Section 3.2 shows that OWL both allows to achieve a higher level of expressivity for representing an ontology of human anatomy, and simplifies the process of building and maintaining complex ontologies by supporting consistency constraints. Section 3.3 shows the expressivity of this anatomy ontology supports the complex reasoning required to infer the consequences of bullet injuries in the region of the heart. Section 3.4 focuses on the comparison of OWL and SWRL respective advantages for optimizing the classification of pacemaker alerts. In all the situations, we showed that if the domain ontologies are available and rich enough, combining them and designing the reasoning portion of the application required a very small amount of work. Unfortunately, we also found repeatedly that such domain ontologies rarely existed.

Chapter 4 presents how classification can be performed when the available informations are incomplete. Section 4.1 is a summary of the open world assumption. Section 4.2 presents a preliminary method for inferring the grade of a tumor according to its description. If the description is incomplete, a classical classification approach may fail because none of the grades requirements are filled. Our method then narrows the range of possible grades by ruling out those incompatible with the information available. Section 4.3 improves the previous method

and proposes a design pattern for modeling clinical trials' eligibility criteria in order to increase patient recruitment.

Chapter 5 presents how ontology can also be used performing semantic similarity-based reasoning. Section 5.1 summarizes the principles of semantic similarity for comparing elements or sets of elements. Section 5.2 proposes a method for computing a generic semantic particularity measure that can be combined with any similarity for a finer interpretation. Section 5.3 presents a method for determining optimal thresholds for semantic similarity and particularity measures.

Chapter 6 presents my research perspectives for producing, querying and analyzing life science data.

Chapter 2

Reasoning based on hierarchies

Outline

Taxonomy-based reasoning is arguably the simplest form of reasoning on an ontology. Nevertheless, this simplicity can also be valuable. It is appropriate whenever the ontology is semantically poor (i.e. a taxonomy or an RDFS hierarchy or polyhierarchy) or when performances are important (i.e. when short answer time possibly over large hierarchies is required). This chapter presents the general principles of taxonomy-based reasoning, and three situations where it was relevant. It demonstrates that even simple reasoning brings added value in situations where using more elaborate tools would be overkill. Section 2.2 emphasizes constraints due to ontology size: it is a use case for generating candidate associations between diseases and pathways using simple ontologies at a time when OWL reasoners could not load them. Section 2.3 is more method-oriented for performing semi-automatic pairing of Web services parameters. Section 2.4 focuses on computation performances of an application providing on the fly semantic expansion of PubMed queries.

2.1 Principle

This section presents why using symbolic data descriptions is a good strategy for analyzing large interdependent datasets, and provides an overview of the associated requirements. We show that the situation we are facing in life sciences is a part of a more general problem. Eventually, we show how RDF and RDFS supports the representation and the analysis of these descriptions.

2.1.1 RDF for describing data

2.1.1.1 Describing data: a generic problem

Annotations as proxies to data Analyzing data can be difficult or time-consuming (usually both), and the problem is even worse if we have to deal with large quantities of data. Moreover, for interdependent data, analyzing some of the data can require the prior analysis of other data. Therefore, saving the result of the interpretation or of the analysis as annotations or metadata is a good strategy so that the next time we need to retrieve some information we do not need to perform the analysis all over again. These annotations can then be used as proxies for faster or more accurate access to results. Naturally, when dealing with large data sets or in order to promote sharing, saving these annotations in a machine-processable format rather than in plain text is desirable.

Data annotation requirements Ideally, these machine-processable data annotations should support the following requirements:

- describe their **nature** (i.e. a binary relation between the data and a set of things sharing some common features): "TGFB1" is a gene, "TGF-β1" is a protein, "apoptosis" is a biological process, "diabetes" is a disease, "The use and misuse of Gene Ontology annotations" is an article,
- describe their **properties** (i.e. a binary relation between a data and some datatype value such as a string, a number, a date, etc.): "TGF-β1" is 390 amino acids-long, "The use and misuse of Gene Ontology annotations" was published in 2008,
- describe **the relations between them** (i.e. a binary relation between two data elements): "TGFB1" is associated to "Homo sapiens", it is located in "chromosome 19" and encodes "TGF-β1", which interacts with the "LTBP1" protein and is involved in "apoptosis", "Seung Yon Rhee" is an author of "The use and misuse of Gene Ontology annotations" which has for subject the "Gene Ontology".
- **combine the descriptions** from different sources either because these sources partially cover the same topic (e.g. metabolic pathways from Reactome and from HumanCYC) or because these sources cover complementary topics (e.g. the genes associated to a disease of interest and the pathways these genes are involved in).

In the previous examples, the resources are typically identified by a string issued by some *de facto* or *de jure* "authoritative source": the human gene "TGFB1" is preferentially referred to by "ENSG00000105329" in Ensembl, "Homo sapiens" by "9606" in the NCBI taxonomy of species, the human proteins "TGF-β1" and "LTBP1" respectively by "P01137" and "Q14766" in Uniprot, the article "The use and misuse of Gene Ontology annotations" by "PMID:18475267" in PubMed, and the Gene Ontology by "`http://purl.bioontology.org/ontology/GO`" in BioPortal. There is an obvious heterogeneity of the identifier patterns among these authorities.

Moreover, what constitutes an "authoritative source" is not always well defined. For example, apoptosis is described among others as a biological process in the Gene Ontology (GO:0006915), as a cellular process in KEGG (ko04210), or as pathway in Reactome (REACT_578). Similarly, "glioma" is identified by "C71" in the 10th version of the International Classification of Diseases ICD10 (but it was "191" in the 9th version), "DI-02566" in Uniprot, "ko05214" in KEGG...

Eventually, some resources may not have been assigned an identifier: as of today, Seung Yon Rhee (the author of PMID:18475267) does not appear to have an ORCID[1] identifier.

Although all the examples we gave are related to life sciences, the problem is more generic.

2.1.1.2 RDF

The Resource Description Framework[2] (RDF) is a W3C recommendation providing a standard model for data interchange on the Web.

Identify resources using IRIs In RDF, a *resource* is anything that can be identified. Identification is performed using Internationalized Resource Identifiers (IRIs), which generalize Uniform Resource Identifiers (URIs) to non-ASCII character sets such as kanji, devanagari, cyrillic... In the remainder of this document we will only use URIs.

[1] `http://orcid.org/`
[2] `http://www.w3.org/RDF/`

URIs syntax follows the pattern:
`<scheme name>:<hierarchical part>[?<query>][#<fragment>]` where `<scheme name>` is typically "`http`" or "`urn`".

Note that although URIs having an `http` scheme name look like URLs, they actually form a superset of URLs as they may not be dereferenced (i.e. they are identifiers, not addresses and there is not necessarily an Internet resource at this address). For example, Uniprot generates URIs for proteins by appending their Uniprot identifier to `http://purl.uniprot.org/uniprot/` and these are URLs (and by transitivity, also URIs and IRIs) so that `http://purl.uniprot.org/uniprot/P01137` is dereferenced either to a Web page or to some RDF description of TGF-β1 depending on the header of the request. Likewise, Gene Ontology generates URIs for its terms by replacing the colon in their identifiers by an underscore, and by appending the result to `http://purl.obolibrary.org/obo/GO_`, but these are not dereferenceable: `http://purl.obolibrary.org/obo/GO_0006915` is an URI that is not an URL.

Also note that URIs specify how to represent a resource identifier but does guarantee uniqueness so that anyone is free to forge as many URIs as wanted to identify something (e.g. bio2rdf uses `http://bio2rdf.org/go:0006915` for referring to GO:0006915 whereas the Gene Ontology uses `http://purl.obolibrary.org/obo/GO_0006915`). Of course, interoperability encourages to reuse existing identifiers whenever possible.

As URIs are cumbersome to deal with by humans, we often use the more convenient prefixed version (e.g. `uniprot:P01137`), but this still requires to specify somewhere that the "`uniprot:`" prefix is actually associated to "`http://purl.uniprot.org/uniprot/`". Prefixed URIs are always unambiguously expanded into full URIs before being processed.

Describe resources with triples In RDF, resources are described using *statements* that are triples composed of a *subject*, a *predicate* and an *object* (noted `<subject> <predicate> <object> .`). The subject is the URI of the described resource. The predicate is the URI of the relation (called a property). The object is a value of the relation for the described resource. This value can be either some URI identifying a resource, or a literal (i.e. a string with an optional indication of a datatype and an optional indication of language). Figure 2.1 on the next page presents two RDF triples sharing the same subject; one of the triples' object is a resource and the other's is a literal. These two triples illustrate how RDF meets respectively the third and the second requirements mentioned in section 2.1.1.1 on page 21. The subject or the object may also be blank nodes, which we will not cover here as it has no impact on RDF expressivity. Note that if the relation can have several values for a resource, this requires as many statements as values.

The fact that some statements can share the same subject (Figure 2.1 on the next page), the same object or that the object of a statement can be the subject of another statement (Figure 2.2 on page 25) result in a directed graph structure connecting resources. As mentioned in the fourth requirement, statements coming from different sources can be combined in a single expanded graph, provided these sources use the same URIs to identify the same things. For example, Figure 2.2 on page 25 combine statements from Uniprot and from Reactome.

RDF specifies a special predicate `rdf:type`[3] for describing the nature of a resource (i.e. a *class* the resource is a member of). For example the statement `uniprot:P01137 rdf:type uniprotCore:Protein` indicates that TGF-β1 (P01137) is an instance of the class `Protein`. This `rdf:type` property allows RDF to address the first requirement.

Figure 2.2 on page 25 also shows how the use of an unique URI across different data sources promotes interoperability and allows to combine complementary descriptions. Here,

[3] `http://www.w3.org/1999/02/22-rdf-syntax-ns#type`

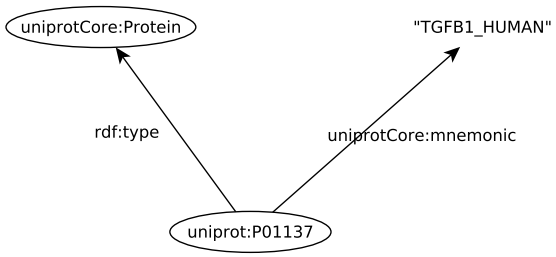

Figure 2.1: Two RDF triples describing the same resource (`uniprot:P01137`). Resources are represented with ellipses, and literals by strings. Properties linking a subject to an object are represented with arrows. URI prefixes are the usual ones (`http://prefix.cc` is your friend).

Uniprot has a triple `uniprot:P01137 rdfs:seeAlso reactome:REACT_120727.4` where the object is the URI corresponding to the "Downregulation of TGF-β receptor signaling" pathway of Reactome, which in turn allows us to retrieve some additional information about this pathway. However, in the same example, Uniprot uses the *`uniprotCore:organism`* property linking to `taxo:9606` whereas Reactome uses the *`biopax3:organism`* property linking to `http://identifiers.org/taxonomy/9606`. This unfortunate use of different properties for representing the species associated to a protein or a pathway, and of different URIs to identify *Homo sapiens* prevents us twice to combine Uniprot and Reactome (e.g. for controlling that Uniprot proteins and the associated Reactome pathways are consistently annotated by the same species or for assisting during this annotation process). Also remember the part about authoritative sources issuing URIs: in this case, both Uniprot and Reactome could have used the URI by the NCBI Taxonomy Database (either the Web page or the corresponding BioPortal resource `http://purl.bioontology.org/ontology/NCBITAXON/9606`). In this particular case, though, Reactome relies on the `identifiers.org` service by the NCBI to provide an additional level of indirection which actually allows to reconcile the Uniprot and the NCBI taxonomy [106]: the `identifiers.org` service lists several URIs associated to `http://identifiers.org/taxonomy/9606`, including `http://purl.bioontology.org/ontology/NCBITAXON/9606` as the primary one, as well as the Uniprot one.

2.1.2 RDFS for describing types

While RDF is adapted for describing resources and relations between resources, RDF Schema[4] (RDFS) provides a vocabulary for describing resources that are classes or predicates. This vocabulary is represented in RDF so that any RDFS statement is also a valid RDF statement.

2.1.2.1 RDFS classes

In RDFS, a *class* is a group of resources, which are its *instances* (the set of the instances of a class is called the extension of the class). Instances and their classes are associated with the `rdf:type` predicate we have seen in the previous section. Classes are themselves resources, so they can be identified by some URI, and described by some properties. Note that two different classes can share the same set of instances (but classes having different sets of instances are necessarily different).

[4] `http://www.w3.org/TR/rdf-schema/`

In RDFS, `rdfs:Class` is the class of all the RDFS classes (and is therefore a metaclass). It is an instance of itself.

RDFS defines the *`rdfs:subClassOf`* property between two classes to represent the fact that the extension of the subject (i.e. the subclass) is a subset of the extension of the object (i.e. the superclass). A hierarchy of subclasses–superclasses is called a taxonomy.

Figure 2.2 shows some examples of associations between instances and their classes using *`rdf:type`* (e.g. between `uniprot:P01137` and `uniprotCore:Protein` for Uniprot or between `reactome:REACT_120727.4` or `reactome:REACT_318.7` and `reactome:Pathway`). It also shows an example of taxonomy using *`rdfs:subClassOf`* between `taxo:9606`, `taxo:9605` and `taxo:207598`. Note that `uniprotCore:Taxon` is a metaclass as its instances are classes.

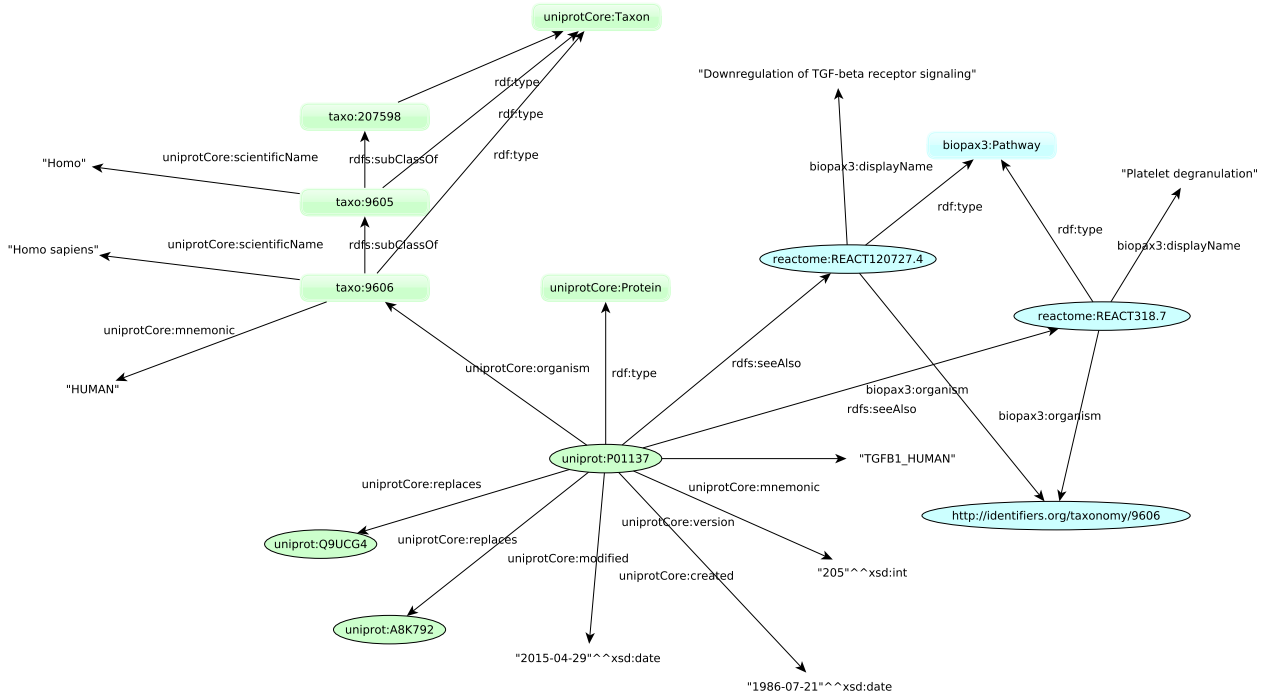

Figure 2.2: Graph of RDF triples describing the same resource (`uniprot:P01137`). Green nodes come from Uniprot and blue nodes from Reactome. Instances and classes are represented by ellipses and boxes respectively. This graph shows typical use of RDF relations between instances or between an instance and a class (*`rdf:type`*), as well as RDFS relations (*`rdfs:subClassOf`* between classes.

2.1.2.2 RDF properties

In RDF, a *property* is a binary relation from one resource (the subject) to another resource (the object). The set of the possible subjects (hence a class) for a property is its domain. The set of the possible objects (hence a class too) for a property is its range. The extension of a property is a subset of the Cartesian product of its domain and its range.

In RDFS, `rdf:Property` is the class of all the RDF properties (i.e. the relations between resources). It is an instance of `rdfs:Class`.

RDFS defines two properties *`rdfs:domain`* and *`rdfs:range`* for defining the domain and the range of RDFS properties (the domain of *`rdfs:domain`* and *`rdfs:range`* is `rdf:Property`

and their range is `rdfs:Class`).

RDFS defines the *`rdfs:subPropertyOf`* property between two properties to represent the fact that the extension of the subject (i.e. the subproperty) is a subset of the extension of the object (i.e. the superproperty). Naturally, declaring that a property is a subproperty of another property implies some additional constraints on their respective domains and ranges.

2.1.3 RDFS entailments

RDF and RDFS support some well-defined entailments which are supported by reasoners and the SPARQL query language. This section provides a simplified overview, please refer to the W3C RDF1.1 semantics[5] for the normative document and particularly to the chapter 9.2[6].

RDFS entailment 1 *The object of an* *`rdf:type`* *property is an* *`rdfs:Class`*:
If *x* `rdf:type` *y* *then* *y* `rdf:type rdfs:Class` .

RDFS entailment 2 *The instances of a class are also instances of its superclass:*
If *x* `rdf:type` *y* *and* *y* `rdfs:subClassOf` *z* *then* *x* `rdf:type` *z* .

RDFS entailment 3 *`rdfs:subClassOf`* *is reflexive:*
If *x* `rdf:type rdfs:Class` *then* *x* `rdfs:subClassOf` *x* .

RDFS entailment 4 *`rdfs:subClassOf`* *is transitive:*
If *x* `rdfs:subClassOf` *y* *and* *y* `rdfs:subClassOf` *z* *then* *x* `rdfs:subClassOf` *z* .

RDFS entailment 5 *The relations of a property also hold for its superproperties:*
If *x* *R1* *y* *and* *r1* `rdfs:subPropertyOf` *r2* *then* *x* *r2* *y* .

RDFS entailment 6 *`rdfs:subPropertyOf`* *is reflexive:*
If *r* `rdf:type rdf:Property` *then* *r* `rdfs:subPropertyOf` *r* .

RDFS entailment 7 *`rdfs:subPropertyOf`* *is transitive:*
If *r1* `rdfs:subPropertyOf` *r2* *and* *r2* `rdfs:subPropertyOf` *r3* *then* *r1* `rdfs:subPropertyOf` *r3* .

Note that by combining RDFS entailments 2 and 4, the instances of a class are also instances of all its ancestors. Similarly, by combining RDFS entailments 5 and 7, a property between a subject and an object can be generalized to all the ancestors of the property.

2.1.4 Typical uses of RDFS entailments in life science

2.1.4.1 Classes hierarchies

Classes hierarchies are the most common structure of ontologies, not only in life sciences. A notable example is Linnaeus' taxonomy of species and the related NCBI taxonomy[7] of all the organisms in the public sequence databases (Figure 2.3 on the facing page). Similarly, life science ontologies from the major repositories OBO Foundry[8] and Bioportal[9] are typically organized as classes hierarchies.

[5] `http://www.w3.org/TR/rdf11-mt/`
[6] `http://www.w3.org/TR/rdf11-mt/#rdfs-entailment`
[7] `http://www.ncbi.nlm.nih.gov/taxonomy`
[8] `http://www.obofoundry.org/`
[9] `http://bioportal.bioontology.org/`

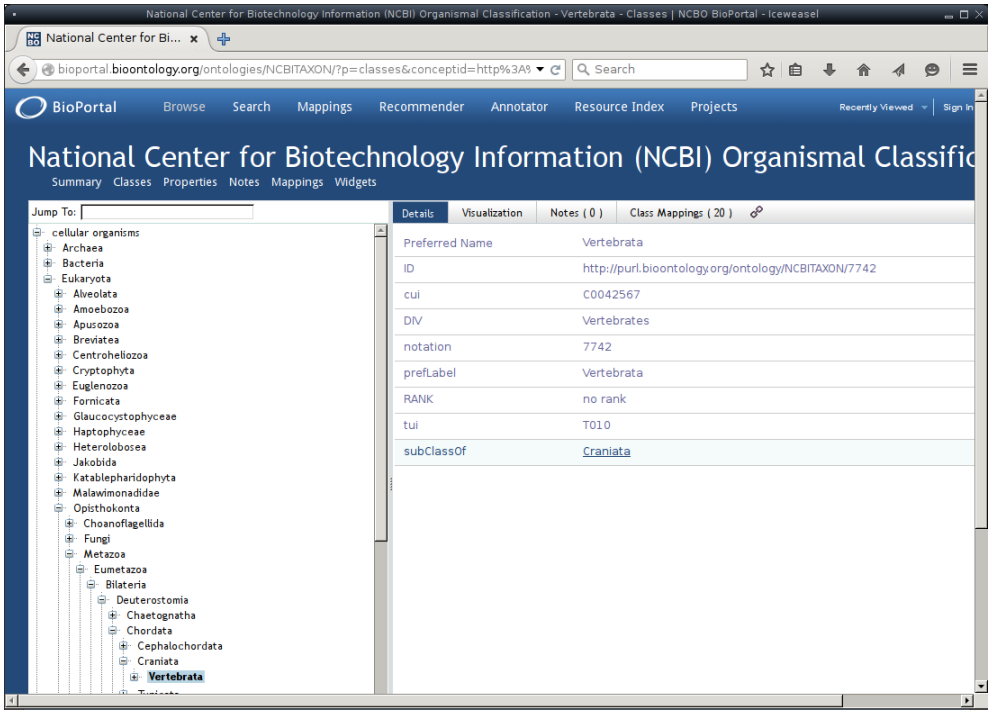

Figure 2.3: The NCBI Taxonomy of species is a (deep) tree-like hierarchy.

Most ontologies are polyhierarchies (i.e. a class can have zero or several direct superclasses such as in Figure 2.4), and few have a tree structure (i.e. all the classes but the root have exactly one superclass such as in Figure 2.3).

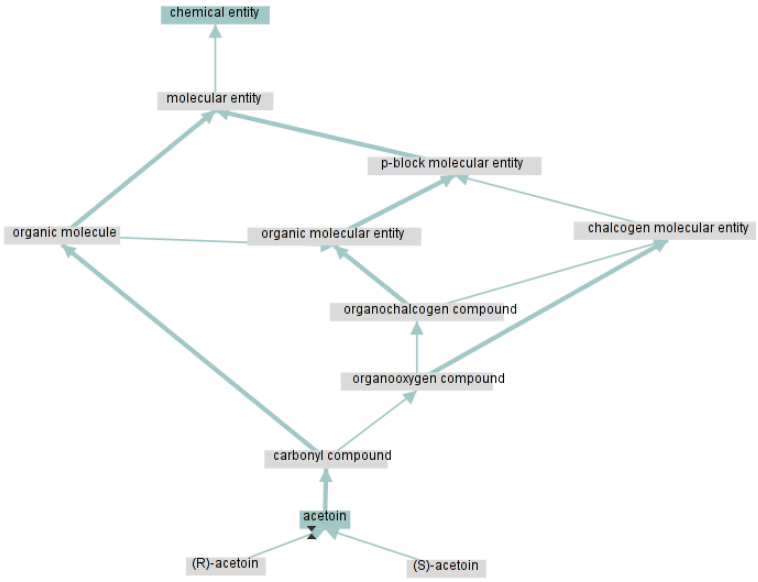

Figure 2.4: The superclasses of acetoin in ChEBI show a polyhierarchy.

Most ontologies have an intricate and deep taxonomy. Some exceptions are "flat" hierarchies,

such as the Online Mendelian Inheritance in Man[10] (OMIM) with a maximal depth of 2, the Enzyme Commission number[11] (EC number) classifies enzymes according to the reactions they catalyze and is organized in 4 levels, or the KEGG Orthology with the first two levels describing pathways categories and the third level pathways (c.f. section 2.2.3.1).

Taxonomy-based reasoning with these ontologies typically involves RDFS entailment rules 2 and 4. Both are used for reconciling the granularity differences between precise annotations and more general queries.

2.1.4.2 Properties hierarchies

Property hierarchies are more seldom used in life science ontologies than classes hierarchies. A typical example is the Gene Ontology (GO) that specifies a *`regulates`* property with two subproperties *`negatively regulates`* and *`positively regulates`*. These three properties are used in a pattern with *`rdfs:subClassOf`*. The *`regulates`* property associates a GO class "`Regulation of X`" with the corresponding GO class "`X`" (using an OWL existential restriction covered in section 3.1.5). The class "`Regulation of X`" has two subclasses "`Positive regulation of X`" and "`Negative regulation of X`", respectively associated to "`X`" by *`negatively regulates`* and *`positively regulates`* (Figure 2.5).

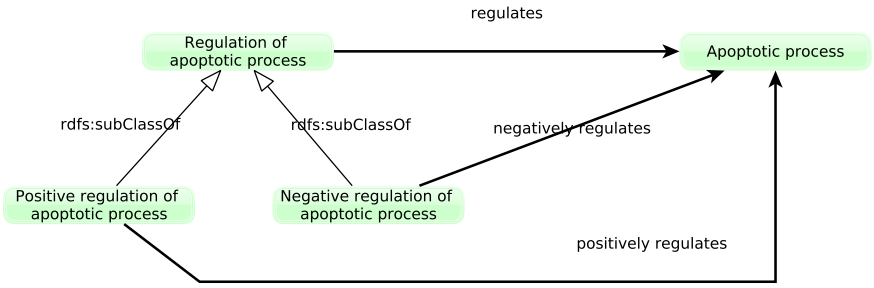

Figure 2.5: Usage of the *`negatively regulates`* and *`positively regulates`* subproperties of *`regulates`* in the Gene Ontology.

Reasoning based on properties hierarchy typically involves RDFS entailment rules 5 and 7. Like classes hierarchies, both are used for handling different levels of precision in the data descriptions. Figure 2.6 on the facing page shows the possible generalizations of "`Positive regulation of leukocyte migrations`" and of '`Leukocyte migrations`" as well as the corresponding regulation relations.

2.1.4.3 Application to annotations

Reasoning based on classes and properties hierarchies is often used for reconciling annotations with different granularities [45]. Because of the definitions of *`rdfs:subClassOf`* and of *`rdfs:subPropertyOf`*, if a data element is annotated by a class, then we can infer that this data element is also annotated by the superclasses. Because of the transitive nature of *`rdfs:subClassOf`* and of *`rdfs:subPropertyOf`*, we can also infer that the data element is also annotated by all the ancestors. In the Gene Ontology, this principle is known as the "True path rule". For example, the gene product `uniprot:P55008` (AIF1, Allograft inflammatory factor

[10] `http://bioportal.bioontology.org/ontologies/OMIM`
[11] `http://www.chem.qmul.ac.uk/iubmb/enzyme/`

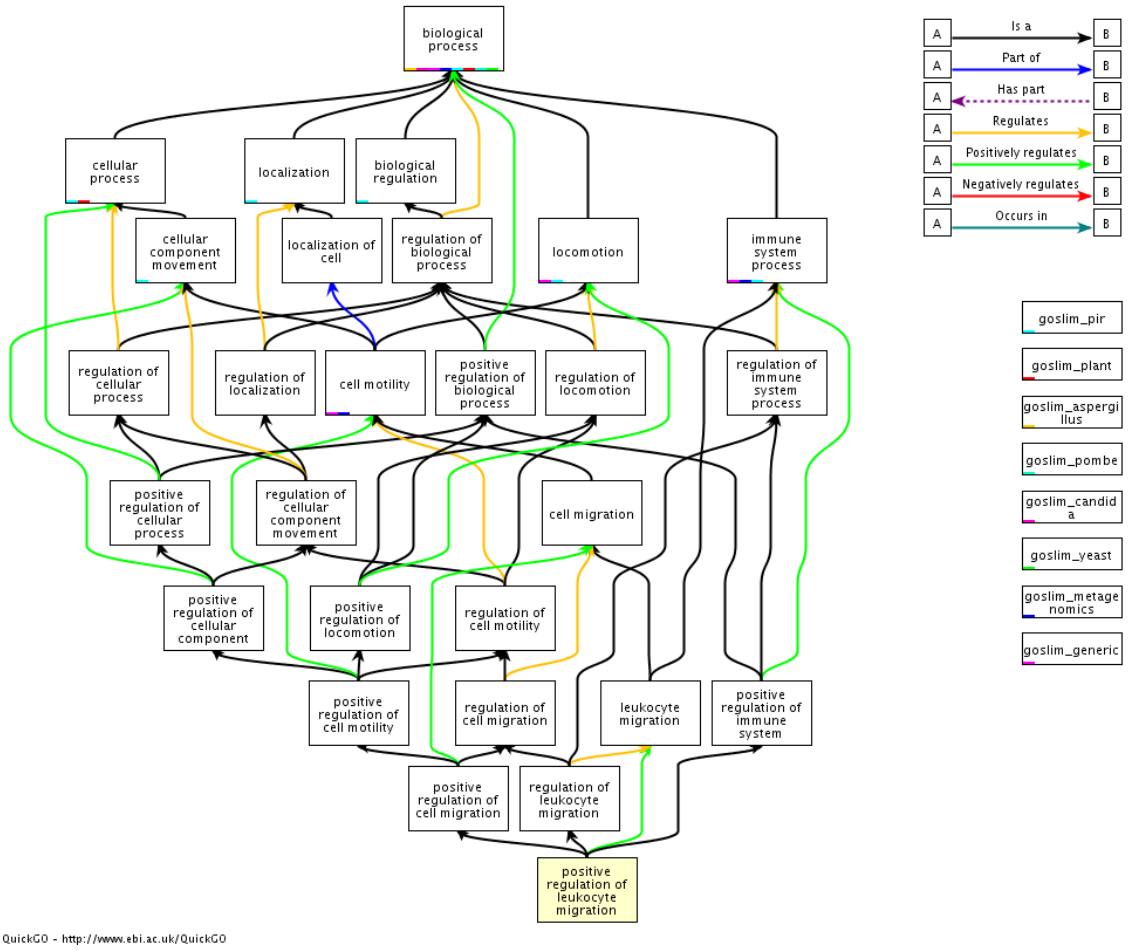

Figure 2.6: Complex mix of *rdfs:subClassOf* hierarchies and of *rdfs:subPropertyOf* hierarchies based on *regulates*, *positively regulates* and *negatively regulates* associating the GO class `Positive regulation of leukocyte migration` and its ancestors to the GO class `Leukocyte migration` and its ancestors (Image by QuickGO `http://www.ebi.ac.uk/QuickGO/`).

1) is annotated (among others) by `GO:0002687` (Positive regulation of leukocyte migration) in *Homo sapiens*. The GO hierarchy (cf. Figure 2.6 on the previous page) allows us to infer that AIF1 is also involved in "cell migration" and in "immune system process". Several articles provide more information on the GO annotations and the related inferences[107, 108]. Livingston et al. also provided an interesting work on the representation of annotations [109].

The "True path rule"-like reasoning is useful in two situations: for analyzing the annotations of data elements, and for querying the data elements annotated by some ontology class. In the first case, we proceed from the data element to its annotations, and in the second case from the annotations to the data elements. To comply with the semantics, reasoning consists in moving up along the hierarchy in the first case, and moving down in the second case.

A typical first case scenario consists in comparing two gene products by analyzing the common GO terms or the ones specific to one of the gene products: comparing their lists of direct annotations is likely to miss some common terms due to the granularity differences, and one should compare the lists of indirect annotations (i.e. the direct annotations and their ancestors).

A typical second case scenario consists in performing some query expansion for retrieving the data elements annotated directly or indirectly by some annotation of interest (e.g. the gene products involved in "immune system process" with GO, or the articles about "infectious diseases" with the MeSH). Retrieving the data elements directly annotated is a trivial database query. However, we should also look for the data elements annotated by some descendant of these annotations, as the true path rule indicates that the annotation of interest is also valid for them.

2.1.5 Synthesis

As we have seen, RDFS-compliant reasoning consists mainly in computing the transitive closure of *`rdfs:subClassOf`* and *`rdfs:subPropertyOf`*. Of course, typical reasoning patterns usually involve combining both. Dedicated RDFS reasoners and query engines have been perfected over the years. The simplicity of the task have allowed them to gain far better performances than *ad-hoc* solutions based on classic programming languages, or relational engines that are notoriously bad at handling transitive closures.

In the remainder of this chapter, section 2.2 shows how an RDF(S) query engine allows to combine multiple ontologies and to query them whereas each of these ontologies was too large to be loaded by an OWL reasoner (even if these ontologies were merely polyhierarchies). This demonstrated that not all tasks on ontologies require an OWL reasoner... and using one can even be counter-productive. Section 2.3 focuses on a reasoning method that fully exploits subclasses and subproperties for pairing Web services parameters. Section 2.4 shows that even for a large ontology such as the Gene Ontology, RDFS reasoning is compatible with on the fly PubMed query enrichment as the time spent enriching the query is negligible compared to the time spent by PubMed for answering the query.

Overall, this chapter shows that RDFS reasoning is valuable even if the medical ontology community was mostly focusing on OWL reasoners (on semantically-simple ontologies).

2.2 Case study: integrating diseases and pathways

This case study focuses on the integration of overlapping ontologies covering different aspects of life science. We created a biomedical ontology associating diseases and pathways using mapping and alignment techniques over KEGG Orthology, Gene Ontology and SNOMED-CT. We represented this ontology in OWL and demonstrated that RDFS queries were expressive enough with acceptable computational performances.
In retrospect, this work is interesting because it identified the need for *a posteriori* resource integration (the linked open data initiative originated around 2007, and ontology alignment and mapping became a very active field in this period), and highlighted the need for using reasoning tools adapted to the task at hand (in those days, DL reasoners could hardly load an ontology, so loading several ontologies was out of question; it would have been overkill anyway because these ontologies are mostly simple taxonomies that hardly use DL features).

This work was a collaboration with Julie Chabalier who was a postdoctoral fellow. It was supported by a grant from Région Bretagne (PRIR) and was originally published in Julie Chabalier, Olivier Dameron, and Anita Burgun. Integrating and querying disease and pathway ontologies: building an OWL model and using RDFS queries. In *Bio-Ontologies Special Interest Group, Intelligent Systems for Molecular Biology conference (ISMB'07)*, 2007 [85].

2.2.1 Context

Use of ontologies within the biomedical domain is currently mainstream (e.g. the Gene Ontology GO [110]). Within a few years, the success of bio-ontologies has resulted in a considerable increase in their number (e.g Open Biological Ontologies[12]). While some of these bio-ontologies contain overlapping information, most of them cover different aspects of life science. However, **an application may require a domain ontology which spans several ontologies**. Rather than to create a new ontology, an alternate approach consists of reusing, combining and augmenting these bio-ontologies in order to cover the specific domain [111].

Associations between classes of genes and diseases as well as associations between pathways and diseases are key components in the characterization of diseases. Different phenotypes may share common pathways and different biological processes may explain the different grades of a given disease. However, this information remains absent in most existing disease ontologies, such as SNOMED CT. Pathway-related information is present in other knowledge sources. The KEGG PATHWAY database is a collection of pathways maps representing our knowledge on the molecular interaction and reaction networks for metabolism and cellular processes [112]. As the GO does not provide direct association with pathways, Mao et al. have proposed to use the KEGG Orthology (KO) as a controlled vocabulary for automated gene annotation and pathway identification [113]. At that time, information about the pathways involved in human diseases has been added to KO.

A major step for addressing this issue is "ontology integration", which sets up relations between concepts belonging to different ontologies. It encompasses several notions: **merging** consists in building a single, coherent ontology from two or more different ontologies covering similar or overlapping domains, **aligning** is achieved by defining the relationships between some of the terms of these ontologies [114] and **mapping** corresponds to identifying similar concepts or relations in different sources [115].

[12] http://obofoundry.org/

The automatic exploitation of the knowledge represented in integrated ontologies requires an explicit and formal representation. Description Logics, and OWL (Web Ontology Language) in particular, offer a compromise between expressivity and computational constraints [116]. However, for leveraging its expressivity, ontologies should contain features such as necessary and sufficient definitions for classes whenever possible, as well as disjointness constraints. While recent works put forward a set of modeling requirements to improve the representation of biomedical knowledge [117, 51], current biomedical ontologies are mostly taxonomic hierarchies with sparse relationships. Even though, dedicated reasoners are hardly able to cope with them.

2.2.2 Objective

The objective of this study was to infer new knowledge about diseases by first integrating biological and medical ontologies and finally querying the resulting biomedical ontology. We hypothesized that most typical queries do not need the full expressivity of OWL and that RDFS is enough for them. In this study, we used the term 'pathway' for metabolic pathways, regulatory pathways and biological processes. The approach presented here consisted in developing a disease ontology using knowledge about pathways as an organizing principle for diseases. We represented this disease ontology in OWL. Following an integration ontology methodology, pathway and disease ontologies have been integrated from three sources: SNOMED CT, KO, and GO. To investigate how information about pathways can serve disease classification purposes, we compared, as a use case, glioma to other neurological diseases, including Alzheimer's disease, and other cancers, including chronic myeloid leukemia.

2.2.3 Linking pathways and diseases using GO, KO and SNOMED-CT

2.2.3.1 KEGG Orthology

The KEGG PATHWAY database was used as the reference database for biochemical pathways. It contains most of the known metabolic pathways and some regulatory pathways. KO is a further extension of the ortholog identifiers, and is structured as a directed acyclic graph (DAG) hierarchy of four flat levels. The top level consists in the following five categories: metabolism, genetic information processing, environmental information processing, cellular processes and human diseases. The second level divides the five functional categories into finer sub-categories. The third level corresponds to the pathway maps, and the fourth level consists in the genes involved in the pathways. The first three levels of this hierarchy were integrated in the disease ontology.

KO hierarchy is provided in HTML format. We extracted the three upper levels of this hierarchy. Each KO class was represented by an OWL class respecting the subsumption hierarchy.

2.2.3.2 Gene Ontology

Gene Ontology is composed of three independent hierarchies representing biological processes (BP), molecular functions (MF) and cellular components (CC). A biological process is an ordered set of events accomplished by one or more ordered assemblies of molecular functions (e.g. cellular physiological process or pyrimidine metabolism). Since we consider all pathways as biological processes, the biological process hierarchy was used to enrich the pathway definitions. The GO BP hierarchy is more detailed than that of KO. It is composed of 27,127 terms spanning 16 levels; for more details about GO structure see [118].

We retrieved the OWL version of GO from the Gene Ontology website[13].

[13] `http://geneontology.org/page/download-ontology`

2.2.3.3 SNOMED CT

SNOMED CT was used as reference source for disease definitions because it is the most comprehensive biomedical terminology recently developed. We used SNOMED to enrich the definitions of human diseases provided by KO.

SNOMED CT is not freely available. However, it is part of the UMLS knowledge Sources [119]. Therefore, we extracted the relevant concepts and their parents, as well as their relations, from the SNOMED CT part of the UMLS. The concepts and relations were respectively represented as OWL classes and properties.

2.2.3.4 Ontology integration

The ontology integration process was based on ontology alignment, which defines relationships between terms, and on ontology mapping, which is a restriction of ontology alignment by taking into account only equivalence relationships between terms.

Figure 2.7 presents an overview of the integration principle. See the original article for details about the method (including the automatic decomposition of KO terms such as "Fructose and mannose metabolism" so that it could be mapped to GO "Fructose metabolic process" and "Mannose metabolic process") and the quantitative results.

The resulting ontology connected diseases from the SNOMED-CT ontology to biological processes from the GO using the *hasPathway* relationship.

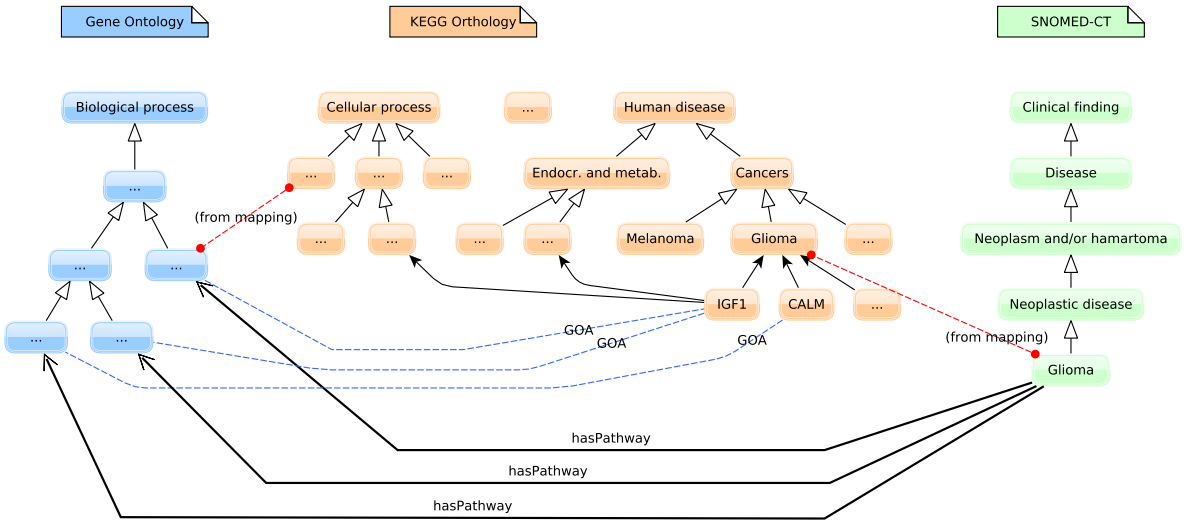

Figure 2.7: Integration of the Gene Ontology (GO), KEGG Orthology (KO) and SNOMED-CT, and the resulting *hasPathway* relations between diseases and pathways. KO is used as a pivot between the GO biological processes for the pathways and diseases from SNOMED-CT.

2.2.4 Querying associated diseases and pathways

Queries can be used either for checking the consistency or for exploiting the resulting integrated bio-ontology.

Typical **consistency queries** consist in detecting if a specific pathway and a more general one are associated with a same disease. Such an imprecision of granularity can either come from

one faulty ontology or from the integration of the knowledge from two ontologies with different granularity.

Typical **queries for exploiting the ontology** involve 1) retrieving the pathways common to several diseases, 2) retrieving the pathways associated with one disease but not with another one, or 3) retrieving the diseases associated with the pathways associated with one class of diseases.

Computing the solutions for both kinds of queries only requires following explicit relations. It does not require OWL-based classification, and can be performed using only the RDFS semantics. RDF repository, and represented the queries using the SeRQL language.

At the time of this study in 2007, we loaded the ontology in a Sesame RDF repository and used SeRQL which was the query language designed by Aduna for the Sesame triplestore [120]. SeRQL advantages over SPARQL (which became a W3C recommendation in 2008) were that it was the query language for the popular Sesame, and that it supported RDFS. Nowadays, SPARQL would be the query language of choice. RDFS support in SPARQL is achieved using property path (e.g. *`rdfs:subClassOf+`* indicates "follow one or more *`rdfs:subClassOf`*") which appeared in 2013 when SPARQL1.1 became a W3C recommendation. We present the SPARQL equivalent of the original SeRQL queries.

2.2.4.1 Redundant disease–pathway associations

When a disease is associated with a pathway, we can infer automatically that it is also associated with all the superclasses subsuming (directly or indirectly) this pathway (Figure 2.8).

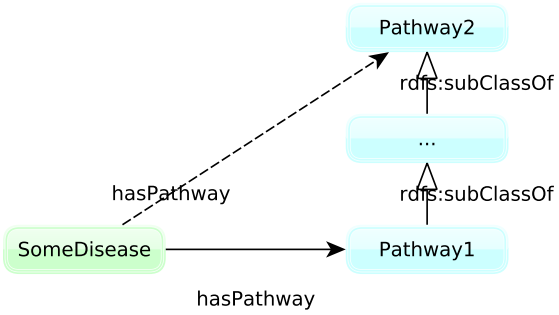

Figure 2.8: The association between `SomeDisease` and `Pathway2` can be inferred using the pathway hierarchy and therefore does not need to be stated explicitly.

The following query retrieves the (disease, pathway) couples linked by a redundant *`has-Pathway`* relation.

```
SELECT DISTINCT ?disease ?redundantPathway
WHERE {
  ?disease dp:hasPathway ?precisePathway .
  ?disease dp:hasPathway ?redundantPathway .
  ?precisePathway rdfs:subClassOf+ ?redundantPathway .
}
```

2.2.4.2 Pathways common to two diseases

The following query retrieves the pathways directly associated to two diseases :

```
1 SELECT DISTINCT ?commonDirectPathway
2 WHERE {
3   ?disease1 dp:hasPathway ?commonDirectPathway .
4   ?disease2 dp:hasPathway ?commonDirectPathway .
5
6   BIND ( snomed:393564001 as ?disease1) # Glioma
7   BIND ( snomed:44054006 as ?disease2) # Type 2 diabetes mellitus
8 }
```

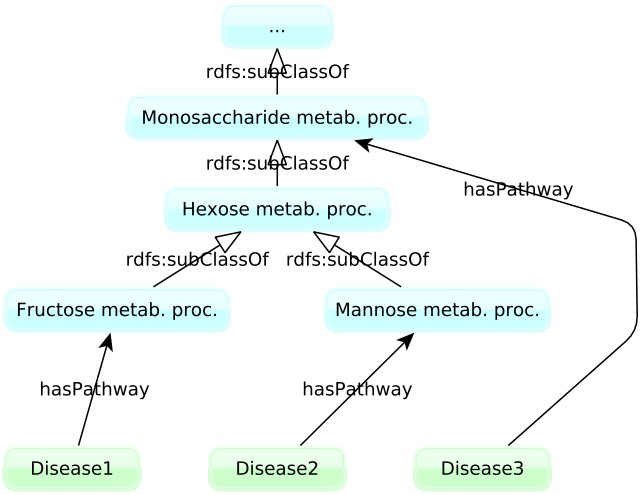

Figure 2.9: None of the three diseases share any direct pathway. However, considering the GO hierarchy allows to recognize that `Disease1` is associated with a pathway more specific than that of `Disease3`, and therefore `Monosaccharide metabolic process` and its ancestors are all common to both diseases (idem for `Disease2` and `Disease3`). Similarly, `Disease1` and `Disease2` are associated with `Hexose metabolic process` and its ancestors.

The previous query is similar to what we could do using a relational database. However, it fails to take the pathway hierarchy into account. If the two diseases are associated with different pathways (e.g. "Fructose metabolic process" for the first and "Mannose metabolic process" for the second) they do not have any direct pathway in common, whereas we could infer from GO that both diseases are associated with the common ancestors of these two terms (in this example "Hexose metabolic process" and its ancestors), as seen in Figure 2.9. The following query retrieves the pathways directly or indirectly associated to two diseases (lines 3 and 4 could have been simplified to `?disease1 dp:hasPathway/rdfs:subClassOf* ?commonPathway` but I kept the verbose version for the sake of clarity; idem for lines 6 and 7):

```
SELECT DISTINCT ?commonPathway
WHERE {
  ?disease1 dp:hasPathway ?pathway1 .
  ?pathway1 rdfs:subClassOf* ?commonPathway .

  ?disease2 dp:hasPathway ?pathway2 .
  ?pathway2 rdfs:subClassOf* ?commonPathway .

  BIND ( snomed:393564001 as ?disease1) # Glioma
  BIND ( snomed:44054006 as ?disease2) # Type 2 diabetes mellitus
}
```

The same principle allows to take the hierarchy of diseases into account in order to retrieve the pathways directly or indirectly associated with a class of diseases. Note that in this case we consider the pathways associated with the disease class or at least one of its subclasses (and not the pathways common to all the diseases of the class). This is mostly because the associations between diseases and pathways are far from being exhaustive, and if one of the diseases is not associated with any pathway, then the entire class would not be either (see chapter 4 for further details about reasoning with incomplete information).

The following query retrieves the pathways directly or indirectly associated with two classes of diseases or one of their subclasses:

```
SELECT DISTINCT ?commonPathway
WHERE {
  ?disease1 rdfs:subClassOf* ?diseaseClass1 .
  ?disease1 dp:hasPathway ?pathway1 .
  ?pathway1 rdfs:subClassOf* ?commonPathway .

  ?disease2 rdfs:subClassOf* ?diseaseClass2 .
  ?disease2 dp:hasPathway ?pathway2 .
  ?pathway2 rdfs:subClassOf* ?commonPathway .

  BIND ( snomed:55342001 as ?diseaseClass1) # Neoplastic disease
  BIND ( snomed:126877002 as ?diseaseClass2) # Disorder of glucose metabolism
}
```

2.2.4.3 Pathways specific to a disease

The following SPARQL query retrieves the pathways directly or indirectly associated with a disease (glioma) but not with another disease (type 2 diabetes mellitus):

```
SELECT DISTINCT ?disease1SpecificPathway
WHERE {
  ?disease1 dp:hasPathway ?pathway1 .
  ?pathway1 rdfs:subClassOf* ?disease1SpecificPathway .

  FILTER NOT EXISTS {
    ?disease2 dp:hasPathway ?pathway2 .
    ?pathway2 rdfs:subClassOf* ?disease1SpecificPathway .
  }

  BIND ( snomed:393564001 as ?disease1) # Glioma
  BIND ( snomed:44054006 as ?disease2) # Type 2 diabetes mellitus
}
```

Note that this query can be modified like we did for the common pathways in order to retrieve the pathways directly or indirectly associated with a class of diseases but not with another class of diseases.

2.2.4.4 Pathways connecting a disease and a class of diseases

The following query retrieves the diseases associated directly or indirectly with the pathways associated with one class of diseases. Note that in this case it is important to consider only the pathways directly associated with the class of diseases or at least on of its subclasses, but that the pathways hierarchy should still be exploited to analyze the pathways associated with `relatedDisease`.

```
SELECT DISTINCT ?relatedDisease
WHERE {
  ?disease rdfs:subClassOf* ?diseaseClass .
  ?disease dp:hasPathway ?pathway .

  ?relatedDisease dp:hasPathway ?pathway2 .
  ?pathway2 rdfs:subClassOf* ?pathway .

  BIND ( snomed:55342001 as ?diseaseClass) # Neoplastic disease
}
```

2.2.4.5 Leukemia, glioma and Alzheimer's disease use-case

For being able to manually check that our queries returned correct results, we considered three diseases: chronic myeloid leukemia, glioma, and Alzheimer' disease. First, we performed some RDFS queries for checking the consistency of the integrated ontology. Among the pathways associated with one disease, 87 are more general than some other pathway associated with this disease (47 for leukemia, 29 for glioma and 10 for Alzheimer's disease). We removed the least specific pathways. We then performed some RDFS queries for comparing diseases by their associated pathways. First, we compared two neurological disorders, namely glioma and Alzheimer's disease (Figure 2.10 on the next page). 8 direct pathway classes involved in glioma were also associated to Alzheimer's disease (86 indirect classes). Then we compared glioma and leukemia; 44 direct pathway classes were shared by these two cancers (165 indirect classes). Finally, 37 pathways are specific to these two cancers (97 indirect classes). Furthermore, the three diseases are associated with pathways themselves associated with glioma.

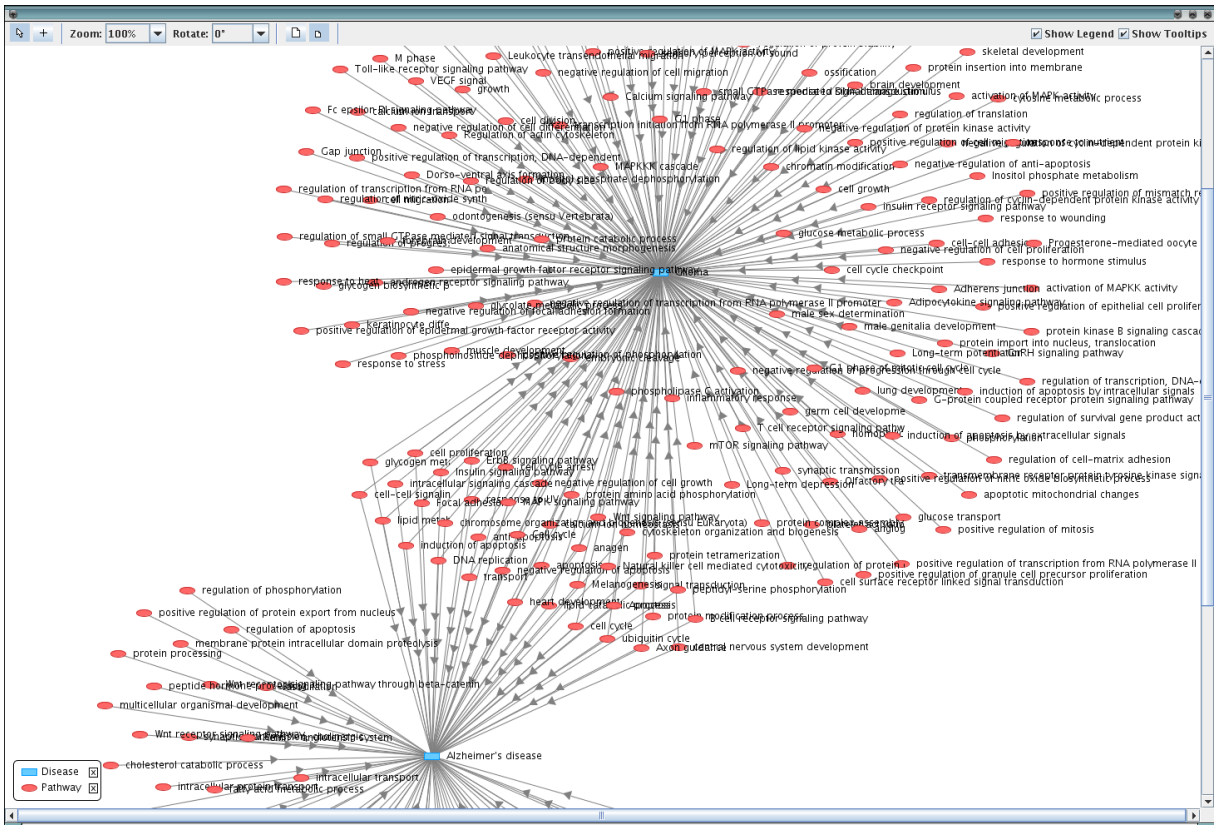

Figure 2.10: Pathways (red ellipses) associated with diseases (blue rectangles): Alzheimer's disease (bottom left) and glioma (center top).

2.3 Methodology: Web services composition

This study focuses on using semantic annotations for helping a user pairing Web services parameters for creating a workflow. From our experience, biologists usually have a rather precise idea of the goal they want to achieve and of the services to use, but they can use some help for the technical aspect of service orchestration. We assumed that the services composing the workflow are known, as well as their relative order. We demonstrated that parameter pairing should not only rely on the type of the parameters (e.g. a string or a date), but also on their nature (e.g. a family name, a city or a creation date as opposed to a validation date). We determined whether pairs of parameters are semantically compatible by examining if they have the same nature or if the output parameter of a service is subsumed by the input parameter of the next service.
In retrospect, this work is interesting because while at that time most of the other works on this domain focused on determining the succession of Web services, we focused on the "next step", i.e. pairing the parameters once this succession is known. Although we made a clear distinction between determining the succession of Web services and pairing the parameters in order to differentiate our work from the others, parameters semantic compatibility could provide some relevant information for guiding the proposition of a succession of Web services.

This work was carried out by Nicolas Lebreton, who I supervised with Anita Burgun. It was originally published in Nicolas Lebreton, Christophe Blanchet, Daniela Barreiro Claro, Julie Chabalier, Anita Burgun, and Olivier Dameron. Verification of parameters semantic compatibility for semi-automatic Web service composition: a generic case study. In *Proceedings of the 12th International Conference on Information Integration and Web-based Applications and Services (iiWAS2010)*, pages 845–848, 2010 [89].

2.3.1 Context

Creating a workflow of Web services is a difficult task [121]. Currently, this is a manual and time-consuming process requiring different expertises. The first step is the selection of the Web services and their arrangement in a sensible order. It requires domain expertise and is typically done by end-users who have an idea of the succession of tasks to perform. Service selection relies on the notion of goal, which is typically represented in task ontologies. The second step consists in pairing the output of a service with one of the inputs of the next one in the workflow. It requires technical expertise for connecting each service input parameters to data or to the output of some other service. Parameter pairing relies on the nature of the parameters, which is typically represented in domain ontologies.

Annotations help automating this tedious process. When present, the WSDL (Web Services Description Language) description of the service is useful, but it only addresses the syntactic level of interoperability (e.g. the input parameter is a *string*). Parameter pairing based on their type (not their nature) is prone to two kinds of errors:

- it misses correct pairings when the parameters are in different formats and would only require a conversion (e.g. the first service output is of type `xsd:date` whereas the second service input is a `xsd:string`);
- it generates incorrect pairings when the parameters are of different natures but represented in the same format (e.g. the first service output is family name of type `xsd:string` and the second service input is a city of type `xsd:string`).

Recognizing the two previous kinds of errors from the correct situations requires some combined reasoning on the type of the parameters (e.g. using XML schemas) and on the nature of the parameters independently from the way they are represented (e.g. a date, a name, a city). OWL-S [122] is the *de facto* standard to represent semantic descriptions of the various Web services. Semantic Web services composition is the act of taking several semantically-annotated Web services and binding them together to meet the needs of the user.

However, semantic description of Web services are scarce and are not really exploited by applications. A well-known tool for multi-domain data analysis is currently Taverna [123] and its workflow language Xscufl (XML Simple Conceptual Unified Flow Language). Taverna allows the user to configure and combine the relevant Web services in a workflow. The Xscufl syntax is used by the Taverna project to store and retrieve workflow definitions. Setting all Taverna's parameters requires a significant knowledge of the Web services and this information is generally not accessible in a software-compatible format that would support automatic assistance. The services combination and the pairing of parameters have to be performed manually, only the execution of the workflow is automated.

2.3.2 Objective

We focused on the use of the semantic annotations to check the compatibility of Web services parameters in a workflow where the order of execution of the Web service is already known.

We demonstrated that OWL-S is expressive enough to support these pairings, and that once the pairings have been decided, an Xscufl file representing the workflow can be automatically generated so that Taverna can execute the workflow.

2.3.3 Semantic compatibility of services parameters

During workflow composition, we assume that the user has already defined the Web services ordering. During each transition from one Web service to the next, we examine all the possible combinations of an output of the first Web service with an input of the next one. Four situations can arise (Figure 2.11 on the facing page):

- *Identical match*: the input and the output have exactly the same kind;
- *Generalization match*: the input is more general than the output of the previous service;
- *Specialization match*: the input is more specific than the output of the previous service. It is up to the user to make sure that in his conditions of use, the first service will always return results that are semantically compatible with the next service input;
- *Incompatibility*: the input cannot be reconciled with the output of the previous service. This is either because the pairing has to be ruled out, or because the ontology is incomplete.

Compatibility is inferred in case of identical match or generalization match.

Compatibility is possible but not guaranteed in case of specialization match. In this case, the type of the output is more general than the type of the next service's input. As the output may not be compatible with the required input type, it is up to the user to determine whether the particular conditions of application guarantee a safe execution.

2.3.4 Algorithm for pairing services parameters

We assumed that the order of the Web services in the workflow is known. We proposed an algorithm for semi-automatically determining which output parameter of a Web service can be

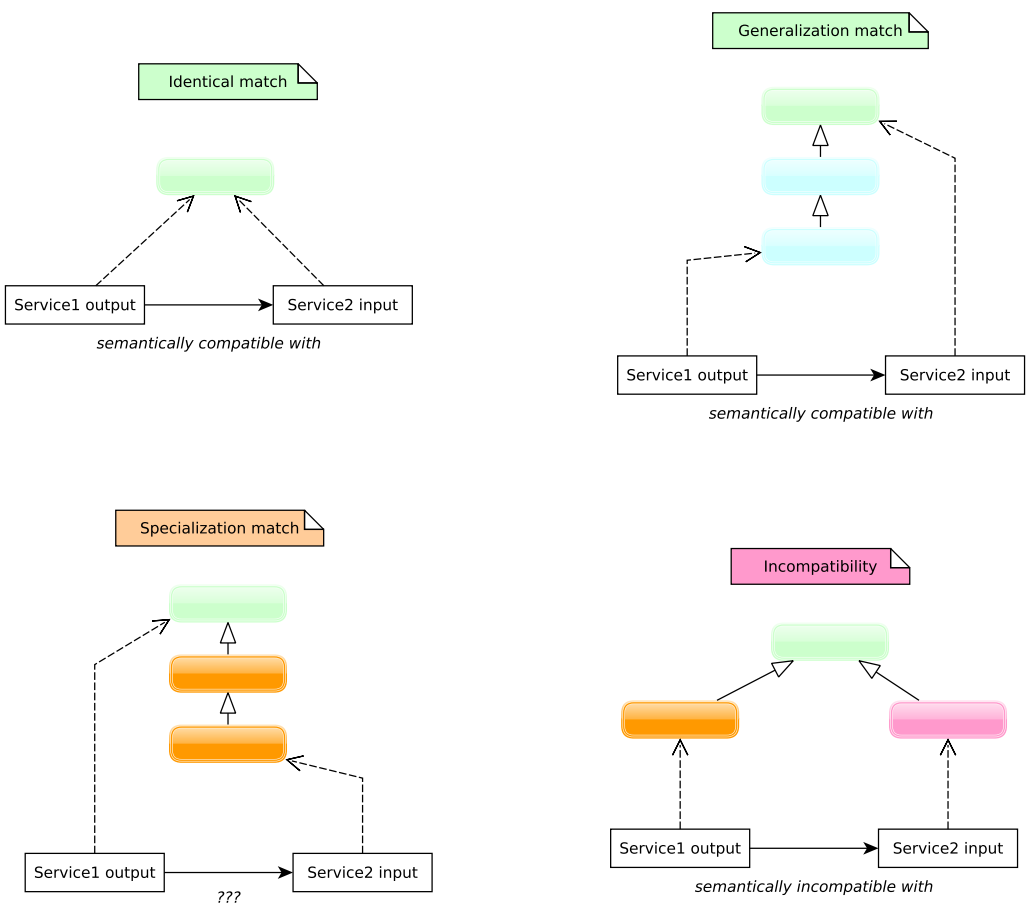

Figure 2.11: Semantic compatibility of a Web service output with the next Web service input depending on the parameters' nature. Compatibility is inferred in case of identical match or generalization match. Compatibility is possible but not guaranteed in case of specialization match. Otherwise, compatibility is ruled out.

paired with which input parameter of the next service. For each input parameter, the algorithm returns a set (possibly empty) of semantically-compatible output parameters, and a set (possibly empty) of potentially-compatible output parameters. Optionally, both sets can be converted to lists if the difference of granularity between the output and the input parameter is considered.

For each pair of connected services, we determine all the pairwise semantic compatibilities of an input parameter of the second service with an output parameter of the first services. Three configurations can arise (Table 2.1 summarizes the possible actions):

- the Web service input is semantically compatible with exactly one output of a previous Web service (either through identical match or through generalization match), and no specialization match was detected. The correct pairing can be validated by the user.

- the Web service input is either semantically compatible with more than one output of a previous Web service (either through identical match or through generalization match), or at least one specialization match was detected. Both lists of candidate pairings are presented to the user who can reject them all or select one ;

- the Web service input is not semantically compatible with any of the outputs of a previous Web service (either through identical match or through generalization match), and no specialization match was detected. The situation cannot be resolved automatically and the user should decide if the problem lies in the workflow itself or in the ontology used to describe the parameters.

		Nb semantically-compatible parameters		
		0	1	>1
Specialization match	0	No compatibility	Validation required	Reject or select one
	>= 1	Reject or select one	Reject or select one	Reject or select one

Table 2.1: Pairing services parameters.

Note that it is important to focus on the input parameters. Parameter pairing consists in choosing one of the outputs of a previous service as the value for the input parameter (and this value is obviously unique, whereas the output of a service can be paired with several inputs of following services). Figure 2.12 on the facing page shows a situation where the input of service S_3 is semantically-compatible with the output of service S_1 and in a specialization match with the output of service S_2. In this case, the user will have to make a choice between the semantically-compatible pair $< O_1, I_3 >$ and a potentially-compatible pair $< O_2, I_3 >$ (the former being the most likely candidate). If we had focused on output parameters, we would have determined that O_1 was most likely paired with I_3, and independently that O_2 was most likely paired with I_3 without realizing that both pairs are concurrent; moreover, in addition to I_3, O_1 could also be paired with the input of another service in addition to S_3.

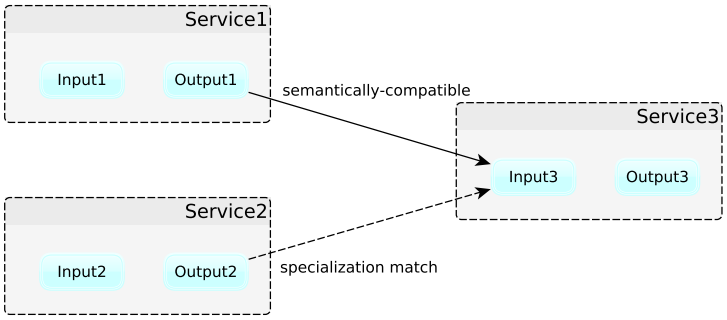

Figure 2.12: Pairing services parameters is performed by considering all the combinations leading to an input parameter and applying Table 2.1 on the facing page.

2.4 Application: ontology-based query expansion with GO2PUB

This application automatically enriches PubMed queries with gene names, symbols and synonyms annotated by a Gene Ontology (GO) term of interest or one of its descendants. GO2PUB is based on a semantic expansion of PubMed queries using the semantic inheritance between GO terms. We demonstrated that this approach yields some relevant articles ignored by the other tools and can be generalized to any biological processes.
In retrospect, this work is interesting because even if the MeSH and the Gene Ontology are arguably the two ontologies that had the greatest impact on the life science community, their integration was original and provided some added value. It should also be noted that the current implementation relies on the GOA databases in a relational format, as they were provided by the EBI. Now that the bio2rdf project released some RDF versions of GOA, the semantic expansion part of GO2PUB could be elegantly rewritten as much simpler SPARQL queries.

This work was initiated and carried out by Charles Bettembourg during his Master's degree internship and then at the beginning of his PhD thesis because he needed to perform PubMed queries on lipid metabolism but found serious limitations with PubMed.

It was originally published in Charles Bettembourg, Christian Diot, Anita Burgun and Olivier Dameron. GO2PUB: Querying PubMed with semantic expansion of gene ontology terms. In *Journal of Biomedical Semantics* 2012, 3:7 [90].

GO2PUB is available at `http://go2pub.genouest.org/`.

2.4.1 Context

The development of high-throughput methods of gene analysis requires to deal with lists of thousands of genes while researchers were used to search the literature only for a few genes at a time. The information retrieval process becomes an increasingly difficult task and needs to be redesigned to provide literature concerning biological problems raised by the gene analyses.

PubMed is the most comprehensive public database of biomedical literature. It comprises more than 21 million entries for biomedical literature from MEDLINE, life science journals, and online books[14]. The typical PubMed user has to read several dozens to hundreds of abstracts

[14] `http://www.ncbi.nlm.nih.gov/pubmed`

to select the relevant ones. More than 4 million articles were added in the last 5 years [15].

A well defined query is important to retrieve as many relevant articles as possible with as few irrelevant ones as possible. Such a query is often more complex than the few loosely-coupled keywords used by most users. There is a need for automatic tools helping the users to build such complex queries that minimize silence and noise [124, 125].

Although PubMed supports MeSH-based query expansion [126], other literature search tools have been developed [127, 128, 129, 130] and evaluated [131]. These can be classified into three major approaches. The first approach, exemplified by tools like SLIM [132], is based on an intuitive interface to set some filters on PubMed queries in order to obtain a better precision than with the basic PubMed querying system. A good proficiency with PubMed *advanced search* brings similar results.

The second approach developed in SEGOPubMed uses a Latent Semantic Analysis (LSA) framework. It is based on a semantic similarity measure between the user query and PubMed abstracts [133]. The authors of SEGOPubMed state that the LSA approach outperforms the other approaches when using well-referenced keywords. Unfortunately, no implementation of SEGOPubMed is currently available. Moreover, this method requires that a corpus of well-referenced keywords be constituted and maintained before the search. Such a corpus is not available (in the biomedical domain) either.

The third approach is based on query enrichment using controlled vocabularies and ontologies. An ontology is a knowledge representation in which concepts are described both by their meaning and their relations to each other [37]. Ontologies are useful to find information relevant to a given topic, particularly through a query expansion process[134]. The automatic handling of the query complexity facilitates query formulation. Expanded queries applied to the Web information retrieval show a systematic improvement over the unexpanded ones [135]. QuExT performs a concept-oriented query expansion to retrieve articles associated with a given list of genes symbols from PubMed and to prioritize them [136].

A frequent goal of gene-related analyses (e.g. transcriptomics) is to identify the genes with different expression across samples analyzed. Thereafter, scientists link their list of genes to more synthetic keywords and functions using Gene Ontology (GO) terms [137] associated to genes thanks to the Gene Ontology Annotation database [138]. At this stage of the gene-related analyses, the keywords to search the literature are not gene names anymore but GO terms. Therefore, tools querying literature with GO terms seem appropriate. GoPubMed [139] uses a text extraction algorithm to mine PubMed abstracts with GO terms. It relies on a local string alignment to compare the GO terms and the abstracts. GoPubMed selects the abstracts containing at least a significant part of the semantic of the GO terms. However, GoPubMed does not follow GO strict rules conveying the semantics of terms. If the annotation of a gene product gp by a Gene Ontology term t is true, then the annotation of gp by any parent of t is equally true [138]. All transitive relation (is a, part of) have to be followed to retrieve these parents. As GoPubMed does not follow this rule, its recall decreases whenever inferences about gene annotations yield new relevant results [140]. None of the existing tools supports a combination of semantics-based and of synonym-based PubMed query enrichment.

2.4.2 Objective

In this study, we hypothesized that the name of the genes annotated by a GO term of interest or one of its descendants can be used as keyword in gene-oriented PubMed queries. The descendants of a GO term are defined according to the Gene Ontology specifications of reasoning about

[15] `http://www.nlm.nih.gov/bsd/licensee/baselinestats.html`

relations. The genes annotated with GO terms are provided by the Gene Ontology Annotation database.

In our system GO2PUB, we propose a new approach that considers not only the genes annotated with a GO term of interest, but also those annotated by a descendant of this GO term, complying with the semantic inheritance properties of GO. GO2PUB's user inputs a list of GO terms of interest, one or more species, and a list of keywords. It generates a PubMed query with the names, symbols and synonyms or aliases of these genes, the species and the keywords and processes PubMed results.

We also performed a qualitative relevance study on our domain of expertise using three queries related to lipid metabolism. In this manuscript, I focus on GO2PUB query expansion. Please refer to [90] for details on the relevance study.

2.4.3 Semantic expansion

Semantic expansion consists in following the semantic inheritance through the GO graph in order to also consider all the descendants of the GO terms specified by the users. Then, the process retrieves the gene names annotated with these terms.

GO2PUB uses these gene names and their synonyms as additional keywords for PubMed queries. Figure 2.13 shows that the expansion identifies five genes associated with the regulation of fatty acid metabolic process, instead of two if the semantic inheritance is ignored.

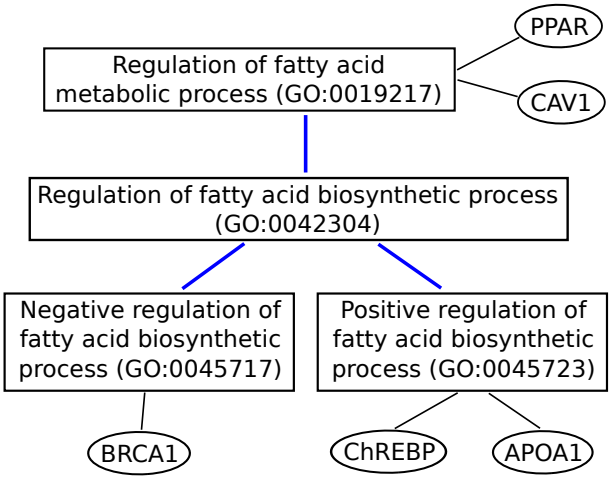

Figure 2.13: Keyword semantic enrichment. For a literature search about the regulation of fatty acid metabolic process, we want to enrich the query with the associated genes. The two genes PPAR and CAV1 are directly annotated by the GO term "Regulation of fatty acid metabolic process" (`GO:0019217`). However, Gene Ontology inheritance properties say that every term inherits the meaning of all its ancestors. Consequently, genes annotated by at least one descendant of the original term (BRCA1, ChREBP and APOA1) should also be considered.

2.4.4 Query generation

GO2PUB creates an expanded PubMed query with the name, symbol and synonyms of genes annotated by one or several GO terms provided by the users, for one or several species. Figure 2.14 on the following page presents the process. The users provide one or several GO terms and species. To further restrict their query, they can also provide as many MeSH terms key-

words as wanted. Furthermore, a "free text" field supports the use of all the other PubMed tags, like [Author], [Journal], etc., and keywords from MeSH terms or free text.

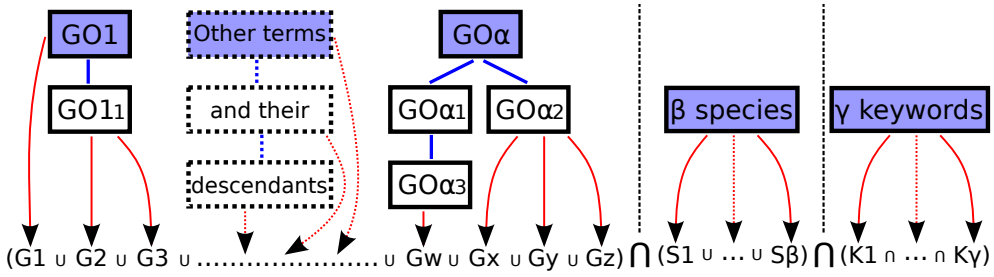

Figure 2.14: GO2PUB Query composition process using the parameters provided by the user. (1) the initial α GO terms (purple boxes) are enriched by their descendants. (2) the genes (here noted G1 to Gz) annotated by the GO terms are retrieved. (3) the query is composed using the names, symbols and synonyms of the genes, the β species (S) and the γ MeSH or free keywords (K).

The first part of each query involves one or more GO terms. The users can enter either the name or the identifier of the GO terms. These terms are suggested when the users start to fill the field. The exact GO term is suggested if the users provide one of its GO synonyms. For example, GO2PUB will search for "lipid biosynthetic process" if the users provide "lipogenesis". When two or more GO terms are entered, GO2PUB makes the union of them ("OR" connector).

Then, the users select one or several species using a name (common or scientific names and their synonyms are allowed) or a NCBI taxon code[16]. In this case, the users can choose to join them (using "OR") or intersect them (using "AND"). Logical connectors "AND" and "OR" are set by default to make the union of species and intersection of keywords, but this can be modified.

Next, the users can enter additional MeSH terms to specify their query. MeSH terms associated to the articles by PubMed are not all of same importance, some of them being classified as "Major topic" (MAJR). We can qualify each keyword as a simple MeSH term or a Major topic. Again, the users can specify the connector between keywords.

GO2PUB retrieves all gene names annotated by each GO term, directly or indirectly through the semantic inheritance properties. It then builds a query on the model "(n gene names, symbols or synonyms separated by OR) AND (m species) AND (p MeSH terms)". The name, symbol and synonyms of each gene compose the first part of the query. They will be searched in title and abstract. Species and keywords chosen by the users make up the second part of the query. Finally, GO2PUB submits to PubMed a query composed of gene names annotated directly or indirectly by the GO terms chosen by the users (name OR symbol OR Synonym), at least one species and some MeSH terms and free keywords. This big query is split into several smaller ones if it exceeds PubMed server URL length limitation. GO2PUB compiles the results and displays all citations numbered and sorted by date.

Figure 2.15 on the next page shows the respective results of PubMed, GoPubMed and GO2PUB for a query about "Lipogenesis in chicken liver". Overall, the comparison of GO2PUB, based on semantic expansion with GoPubMed, based on text mining techniques showed that both tools are complementary.

[16] `http://www.ncbi.nlm.nih.gov/Taxonomy`

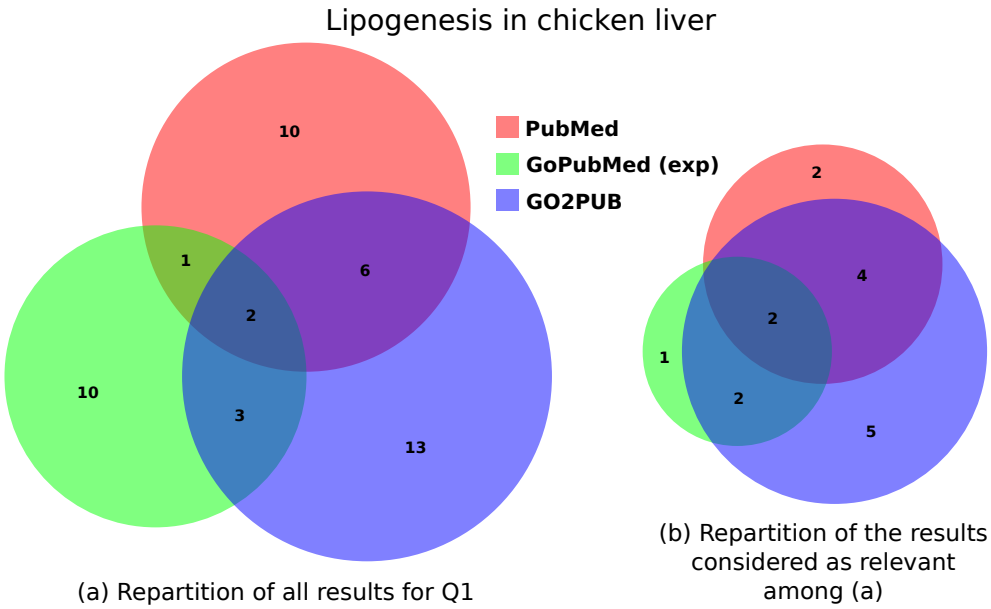

Figure 2.15: Comparison of the PubMed, GoPubMed and GO2PUB results for query "Lipogenesis in chicken liver". (a) displays the repartition and intersections of these results. (b) displays the repartition and intersections of the results considered as relevant.

2.5 Synthesis

What the three studies have in common All three studies require simple taxonomy-based reasoning on OWL ontologies. However, in these cases, this simplicity was well adapted to the situation.

It should be noted that in the first half of the 2000's, OWL reasoners were perceived as the preferred solution for querying OWL ontologies even if (1) most ontologies were essentially RDFS taxonomies represented in OWL, (2) the reasoners could hardly manage to load medium-size ontologies and (3) classification was extremely slow if it successfully terminated (even considering the ontologies simplicities). This can be explained by the fact that unfortunately, few connections existed within the semantic Web community between the people focusing on biomedical ontologies who were mostly DL-oriented, and the people developing RDF and SPARQL. Open source ontology editors such as OILed and Protégé have been coupled with DL reasoners since the early 2000's, but have not supported simple and efficient connections with query languages from the triplestore community until recently (contrary to TopBraid composer). Nowadays, ontologies have mostly evolved in size but not in semantic complexity [118].

Therefore, even if the topics of the first two studies seem outdated now, the principle is still all the more relevant that major life science data sources such as Uniprot[17] or the EBI[18] are providing their data as RDF and offer dedicated SPARQL endpoints, and initiatives such as bio2rdf[61] and the NCBO BioPortal[49, 141, 142] provide an integrated access in RDF to life science datasets and ontologies. Other initiatives such as purl[19] and identifiers.org[20] provide solutions for perennial and location-independent identifiers [106, 64].

[17] `http://beta.sparql.uniprot.org/`
[18] `https://www.ebi.ac.uk/rdf/`
[19] `htts://purl.org`
[20] `http://identifiers.org/`

Integration of diseases and pathways Using SNOMED-CT and the Gene Ontology for integrating diseases and pathways combined the hierarchies of these two ontologies. It was then possible to identify pathways or families of pathways shared by two diseases even if these diseases are directly associated to different pathways, provided the pathways have some common ancestor. Conversely, two pathways may be associated to different but similar diseases.

Current analysis techniques go beyond the simple detection of shared common elements and focus on whether the number of shared elements is greater than what we could expect by random selection [143].

If all the resources had existed in RDFS or OWL, the current version of this work would have consisted in producing an OWL version of the mappings between SNOMED-CT diseases and GO biological processes, and in performing SPARQL federated queries over SNOMED-CT, GO and the mappings.

Composition of Web services We proposed a simple generic algorithm compatible with any annotation framework such as WSMO, OWL-S or SAWSDL. The major limitations for its deployment were the lack of appropriate domain ontology, as well as (understandably given the lack of ontologies) the lack of semantic annotations for Web services and particularly for their parameters [144].

Our semantic compatibility algorithm was only tested on linear workflow. Its generalization to more complex control structures remains to be studied.

Semantic query expansion with GO2PUB GO2PUB performs a semantic expansion of the GO terms of interest complying with the semantic inheritance through the GO graph before retrieving the corresponding genes to enrich the query. Using the semantic inheritance properties of the GO graph was useful, as the more descendants a GO term has, the more relevant results GO2PUB yields.

GoPubMed does not follow the semantic inheritance properties of GO. We manually expanded GoPubMed queries and compared it to GO2PUB. This showed that query expansion would be a valuable extension for GoPubMed.

Both GO2PUB and GoPubMed retrieved relevant articles ignored by PubMed. As most of the results obtained by GO2PUB and GoPubMed are relevant in the qualitative study and in the generalization study, the intersection of GoPubMed and GO2PUB results decreases noise. As each tool yields relevant articles ignored by the other, the union of their results also decreases silence.

GO2PUB seems less suited for queries involving either general GO terms or GO terms with few or no descendants. Indeed, with general GO terms, GO2PUB considers a lot of descendants, and therefore a lot of genes. We expect this to increase the noise as some of the genes will be irrelevant. Conversely, GO terms having few or no descendants are associated with few genes. We do not expect semantic expansion to benefit these highly specific queries yielding only a few PubMed results.

Overall, GO2PUB performed better than GoPubMed and PubMed. GO2PUB brought relevant results ignored by GoPubMed even when adding a manual query expansion for GoPubMed. Conversely GoPubMed text mining approach found relevant articles ignored by GO2PUB. This demonstrates GO2PUB relevance and its complementarity with GoPubMed.

What we learned

- Ontologies are getting bigger much more than they are becoming semantically richer (see [118] for the Gene Ontology). Even though, DL-based reasoners can hardly keep up with the size, so reasoning on rich ontologies is still prohibitive nowadays.

- OWL-based reasoning is not necessary in all situations.
- SPARQL is increasingly well adapted in terms of expressivity and of capability of dealing with large quantities of information (the linked data initiative is soaring).
- The fact that SPARQL is well adapted is a good news in the short term as it obviously fills a need but it may also be a bad news in the longer term as it may turn people away from the endeavor of producing semantically-richer ontologies (which may or may not be a bad thing).

Chapter 3

Reasoning based on classification

Outline

RDFS requires to explicitly state that a resource is an instance of a class or that a class is a subclass of another class. However, we should be able to recognize both situations from the class characteristics. This then brings the need to formalize class characteristics. This formal representation of class characteristics can even be used during the the ontology maintenance process for ensuring completeness and internal consistency (like some non-regression tests) [145]. I proposed a rule-based solution for maintaining the internal consistency of the brain cortex anatomy ontology I developed during my PhD [69]; as this was before the availability of OWL, it was not included in this manuscript.

Description Logics (DL) provides semantically-rich constructs such as disjointness, existential and universal restrictions as well as necessary and sufficient definitions necessary for performing inferences beyond simple taxonomy-based reasoning [38, 71, 116, 146, 147, 51, 43, 54].

All these reasoning capabilities require that ontologies actually implement these various features, which was not the case in the early days of OWL. Section 3.2 shows the difficulty of generating an OWL version of the ontology of human anatomy from a frame-based representation (which fortunately for us was the result of an elaborate and rigorous modeling effort by Cornelius Rosse and his team).

Section 3.3 shows that a portion of this OWL version of the human anatomy ontology was instrumental in the Virtual Soldier project for inferring the consequences of bullet injuries in the region of the heart.

Section 3.4 focuses on the comparison of OWL and SWRL-based reasoning for optimizing the classification of pacemaker alerts.

3.1 Principle

Description Logics (DL) are formal languages allowing to represent characteristics for sets of objects [146, 147]. They were created to remedy semantic networks and frame languages lack of formally-defined semantics [148]. There are several DL languages with different expressivity. This section focuses on the OWL language, which was developed in the Semantic Web context. OWL1.0-DL was based on $\mathcal{SHOIN}^{(\mathcal{D})}$ [116] (OWL1.0 also defined OWL-Full, which did not belong to the DL family in order to support metaclasses ; except when specified otherwise, we focus on DL). Its limitations led to the development of OWL2.0, which extends OWL1.0-DL and is based on $\mathcal{SROIQ}^{(\mathcal{D})}$ [149]. This section presents an overview of OWL main characteristics

that were used for biomedical data analysis. It is a summary of "OWL2 direct semantics"[1] and "OWL2 RDF-based semantics"[2].

3.1.1 OWL Classes

DL represent "concepts" or "classes" as sets of individuals defined in intension. The function that associates a class and a set of individuals is the *interpretation function*, noted $\mathcal{I}$.

Δ is the set of all individuals for a particular domain. For a set of classes, there can be many Δ, and therefore many interpretation functions. Neither Δ nor $\mathcal{I}$ are usually completely known. As the notion of genericity is a key aspect of ontologies, ontologies specify characteristics of classes and relations between them that are valid for all sets of individuals and all interpretation functions.

Note that two classes can be associated to the same set of individuals by an interpretation function $\mathcal{I}_1$ in a particular domain, but to different sets by an interpretation function $\mathcal{I}_2$ in another domain. Classes that are associated to the same set of individuals for all the interpretation functions are equivalent (but each retains its identity). Similarly, a class can be associated to an empty set of individuals according to $\mathcal{I}_1$ but not according to $\mathcal{I}_2$. A class that is associated to an empty set of individuals for all interpretation functions is unsatisfiable or inconsistent.

Having a set-based definition for DL classes allows to define a semantics based on set operations on their interpretation. This formally-defined semantics is useful because it supports reasoning. On counterpart, this means that we will focuses on the classes that comply with this semantics. Because `rdfs:Class` was loosely defined as the range of the ***`rdf:type`*** property, OWL introduces the notion of `owl:Class` (note how prefixes are useful for distinguishing the two) for designating classes that are sets of individuals (so metaclasses are not allowed, contrary to RDFS).

```
owl:Class rdfs:subClassOf rdfs:Class
```

OWL also defines two special classes: `owl:Thing` (also noted top or $\top$) and `owl:Nothing` (also noted bottom or $\bot$).

```
owl:Thing rdf:type owl:Class
```

$$\top^{\mathcal{I}} = \Delta$$

```
owl:Nothing rdf:type owl:Class
```

$$\bot^{\mathcal{I}} = \varnothing$$

Of course, the RDFS definition of ***`rdfs:subClassOf`*** (noted $\sqsubseteq$ between classes) remains valid for OWL classes. In this case, for all interpretation function, the set of instances of the subclass is a subset of the set of instances of the superclass. This is used in ontologies for representing the characteristics shared by all the instances of the class by making the class a subclass of some logical expression combining the instances of other classes.

$$\begin{cases} \texttt{C1 rdfs:subClassOf C2} \\ \Leftrightarrow \\ C1^{\mathcal{I}} \subseteq C2^{\mathcal{I}} \end{cases}$$

[1] `http://www.w3.org/TR/owl2-direct-semantics/`
[2] `http://www.w3.org/TR/owl2-rdf-based-semantics/`

OWL also provides an *owl:equivalentClass* property (noted ≡ between classes) to specify that for all interpretation function, the two classes have the same set of instances. This is used extensively in (semantically-rich) ontologies to provide at least one necessary and sufficient definition for a class by making it equivalent to some logical expression combining the instances of other classes. A class with a necessary and sufficient definition is called a "defined class" by opposition to "primitive classes".

$$\left\{\begin{array}{l} \texttt{C1 owl:equivalentClass C2} \\ \Leftrightarrow \\ C1^{\mathcal{I}} = C2^{\mathcal{I}} \end{array}\right.$$

3.1.2 Union and intersection of classes

The union (resp. intersection) of two classes, noted ⊔ (resp. ⊓) is a class which set of instances is the union (resp. intersection) of the sets of instances of the two classes.

$$\left\{\begin{array}{l} \texttt{owl:unionOf(C1, C2) rdfs:subClassOf owl:Class} \\ \Leftrightarrow \\ (C1 \sqcup C2)^{\mathcal{I}} = (C1^{\mathcal{I}} \cup C2^{\mathcal{I}}) \end{array}\right.$$

$$\left\{\begin{array}{l} \texttt{owl:intersectionOf(C1, C2) rdfs:subClassOf owl:Class} \\ \Leftrightarrow \\ (C1 \sqcap C2)^{\mathcal{I}} = (C1^{\mathcal{I}} \cap C2^{\mathcal{I}}) \end{array}\right.$$

For example, this can be used to define the class `Finger` by making it an equivalent class to the union of `Thumb`, `Index`, etc. Similarly, we can define the class `LeftThumb` as the intersection of `Thumb` and `LeftSideOrgan`.

3.1.3 Disjoint classes

OWL relies on the open world assumption (cf. Section 4.1), so by default, two classes can share some instance with an interpretation function, and not have any instance in common with another interpretation function. The *disjointWith* property specifies that two classes can never have any instance in common. When designing ontologies, sibling classes are usually disjoint (e.g. the five subclasses of `Finger`: `Thumb`, , `MiddleFinger`, `RingFinger` and `LittleFinger` are pairwise disjoint) but not always (e.g. if we also consider the two other subclasses `LeftFinger` and `RightFinger`, which are also pairwise disjoint but should not be assumed to be disjoint from the five others).

Stating explicitly that two classes are disjoint is useful both for avoiding logical inferences that would not be consistent with the domain knowledge and for ruling out options in case of partially-known information. Several examples are presented in Chapter 4

$$\left\{\begin{array}{l} \texttt{C1 owl:disjointWith C2} \\ \Leftrightarrow \\ (C1^{\mathcal{I}} \cap C2^{\mathcal{I}}) = \varnothing \end{array}\right.$$

3.1.4 Negation: complement of a class

The negation of a class (noted ¬) is a class which set of instances is the complement of the set of instances of the original class (i.e. the set of individuals that are not instances of the class). By definition, a class and its negation are disjoint.

$$\begin{cases} \texttt{owl:complementOf(C) rdfs:subClassOf owl:Class} \\ \Leftrightarrow \\ (\neg C)^{\mathcal{I}} = (\Delta \setminus C^{\mathcal{I}}) \end{cases}$$

For example, the intersection of the class `Patient` and of the negation of its subclass `DiabeticPatient` represents the non-diabetic patients (we will see in section 3.1.5 how to define `DiabeticPatient` as the patients suffering from diabetes)

3.1.5 Existential and universal restrictions

An existential restriction (noted ∃ ***property*** . `Class`) is the class which instances are the individuals related to at least an instance of `Class` (direct or indirect) through ***property*** (or a subproperty). For example ∃ ***suffersFrom*** `Diabetes` is the set of the individuals being the subject of a triple which predicate is ***suffersFrom*** and which object is an instance of `Diabetes`.

$$(\exists \textit{property}.\texttt{Class})^{\mathcal{I}} = \{i \in \Delta | \exists j \in \texttt{Class}^{\mathcal{I}}, (\textit{property}^{\mathcal{I}})(i,j)\}$$

An universal restriction (noted ∀ ***property*** . `Class`) is the class which instances are the individuals for which ***property*** (or a subproperty) only leads to instances of `Class` (direct or indirect), i.e. the individuals not related through ***property*** to anything that is not an instance of `Class`. Individuals not related to anything through ***property*** also qualify as instances of the universal restriction. For example, ∀ ***hasMedication*** `Anticoagulant` is the set of individuals whose only medications are anticoagulant, including those who do not have any medication.

$$(\forall \textit{property}.\texttt{Class})^{\mathcal{I}} = \{i \in \Delta | \forall j \in \Delta, (\textit{property}^{\mathcal{I}})(i,j) \Rightarrow j \in \texttt{Class}^{\mathcal{I}}\}$$

Existential and universal restrictions are typically used:

- as superclasses for representing characteristics shared by all the instances of a class. In this case, the reasoning is "if you are an instance of the class, then you match the condition". For example, all the ventricular cavities are filled with blood (and other anatomical structures can also be filled with blood) `VentricularCavity` ⊏ (∃ ***filledWith*** . `Blood`). If you are an instance of `VentricularCavity`, then you are related to at least one instance of `Blood` through ***filledWith***. Even if the relation is not explicitly stated, we know that there must be at least one.

- as equivalent classes for representing characteristics that define a class. In this case, the reasoning can be either "if you are an instance of the class, then you match the condition" or "if you match the condition, then you are an instance of the class". For example, a diabetic patient is an individual who suffers from diabetes `DiabeticPatient` ≡ (∃ ***suffersFrom*** . `Diabetes`).

3.1.6 Cardinality restrictions

A minimum cardinality restriction (noted *min* x ***property*** . `Class`) is a class which instances are the individuals related to at least x instances of `Class` through ***property***.

A maximum cardinality restriction (noted *max* x ***property*** . `Class`) is a class which instances are the individuals related to at most x instances of `Class` through ***property***.

An exact cardinality restriction (noted *exactly* x ***property*** . `Class`) is a class which instances are the individuals related to exactly x instances of `Class` through ***property***.

$$(min\ x\ \textit{property}.\texttt{Class})^{\mathcal{I}} = \{i \in \Delta \mid card(j \in \texttt{Class}^{\mathcal{I}} | \textit{property}^{\mathcal{I}}(i,j)) \geq x\}$$

$$(max\ x\ \textit{property}.\texttt{Class})^{\mathcal{I}} = \{i \in \Delta \mid card(j \in \texttt{Class}^{\mathcal{I}} | \textit{property}^{\mathcal{I}}(i,j)) \leq x\}$$

$$(exactly\ x\ \textit{property}.\texttt{Class})^{\mathcal{I}} = \{i \in \Delta \mid card(j \in \texttt{Class}^{\mathcal{I}} | \textit{property}^{\mathcal{I}}(i,j)) = x\}$$

Because of the open world assumption, cardinality restrictions are a good complement to existential and universal constraints. For the class `Hand`, one could use six existential constraints on ***hasPart*** to specify that a hand has a palm, a thumb, an index, etc. This would still allow an instance of `Hand` to have another part that would be an instance of `Lung`. For preventing this, one could add a closure axiom to ***hasPart***, saying that `Hand` is a subclass of the the things having only parts in the union of `Palm`, `Thumb`, `Index`, etc. However, it would still be acceptable to have an instance of hand with seven palms, three thumbs and two of each other fingers. The right solution here would be to replace the six existential constraints by six exact cardinality constraints and to retain the closure.

3.1.7 Property chains

A property chain (noted ***R1*** o ***R2***) is a property formed by following ***R1*** and ***R2***.

$$(\textit{R1}\ o\textit{R2})^{\mathcal{I}} = \{(i,j) \in \Delta^2 \mid \exists k \in \Delta, (i,k) \in (\textit{R1})^{\mathcal{I}} \wedge (k,j) \in (\textit{R2})^{\mathcal{I}}\}$$

For example, property chains are used in the Gene Ontology (not otherwise known for its semantic richness) for modeling inference[3]. Figure 3.1 on the next page shows the inference pattern "if A positively regulates B and B is part of C, then A positively regulates C". It uses a property chain for modeling the condition and ***rdfs:subPropertyOf*** for modeling the inference.

3.1.8 Synthesis

As we have seen, Description Logics provide the formal foundation for explicitly representing constraints and definitions about classes, as well as relations between classes.

In the remainder of this chapter, section 3.2 shows how rich these formal descriptions can be. It follows that building a semantically-rich ontology can be difficult, but conversely that adopting design patterns can (and should) also simplify both design and maintenance. Section 3.3 shows that once semantically-rich ontologies are available, complex reasoning can be incorporated into an application with minimal development. Finally, section 3.4 compares several modeling strategies in term of complexity and performance.

[3] `http://geneontology.org/page/ontology-relations`

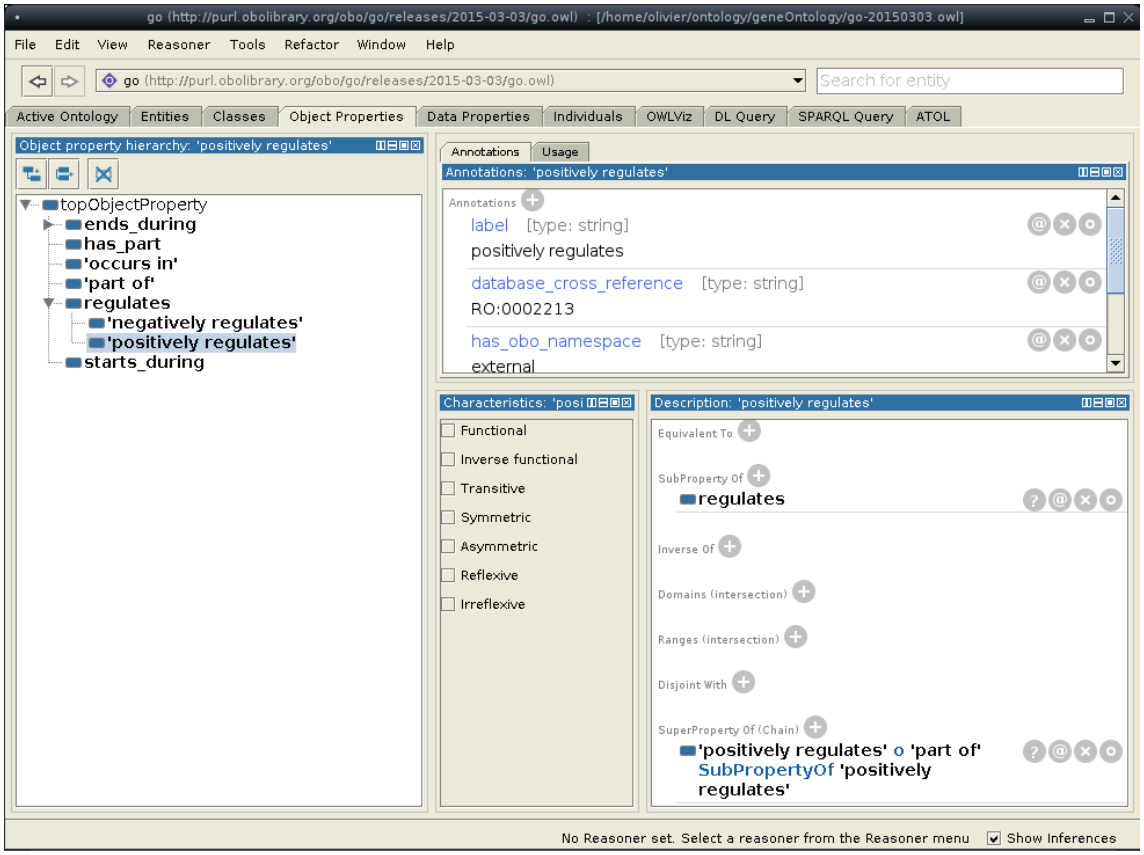

Figure 3.1: Property chain in the Gene Ontology: if x *positively regulates* y and y *part of* z, then we can infer that x *positively regulates* z.

3.2 Methodology: Description-logics representation of anatomy

This study focuses on converting the frame-based representation of the ontology of normal human anatomy (the FMA) into OWL, which is more adapted to the reasoning task we envisioned (c.f. section 3.3). The challenges were (1) preserving and possibly extending the semantic richness of the original ontology (contrasting with the other ontologies we have seen previously that were mostly taxonomies) and (2) dealing with the large size of the ontology, which placed it out of reach for the reasoners. We showed that perfect conversion was impossible: we had to forgo some the the frame-specific aspects, but on the other hand we could add some OWL-specific ones such as disjointness or coverage. We recognized that the computational constraints required some simplification for the resulting ontology to be usable by applications, but that these simplifications were application-specific. We proposed a solution that delivers an OWL-Full representation as expressive as possible, and a "Virtual FMA-OWL" mechanism (detailed in the original article) providing access to the FMA on a concept-by-concept basis for further simplifications and adaptations.
In retrospect, this work is interesting because it showed that even if Description Logics offer some highly desirable knowledge modeling primitives, using these primitives consistently and systematically during ontology design is difficult and in turn requires some automation. At that time I mostly used dedicated *ad hoc* scripts on the FMA [76] or on the ontology of brain cortex anatomy I developed during my PhD [69], but this problem was later addressed by others in more principled approaches such as ontology design patterns[43] or more recently the SPIN[a] and Shape Expression[b] (ShEx) languages for expressing constraints on RDF(S) graphs.

[a] `http://spinrdf.org/`
[b] `http://www.w3.org/2001/sw/wiki/ShEx`

This work was originally published in: Olivier Dameron, Daniel L. Rubin, and Mark A. Musen. Challenges in converting frame-based ontology into OWL: the Foundational Model of Anatomy case-study. In *American Medical Informatics Association Conference AMIA05*, pages 181–185, 2005 [75] and later extended in: Olivier Dameron and Julie Chabalier. Automatic generation of consistency constraints for an OWL representation of the FMA. In *10th International Protégé Conference*, 2007 [76].

3.2.1 Context

In the medical domain, anatomy is a fundamental discipline that underlies most medical fields [150]. The Foundational Model of Anatomy (FMA) is the most complete ontology of canonical (i.e. healthy) human anatomy [151]. It strictly follows a principled modeling approach and included more than 70,000 concepts and 1.5 million relationships at the time of the study in 2004.

The FMA is represented in a frame language [152]. However, for some applications it is desirable to use an OWL representation of the FMA, either for reasoning purposes [73] or for integrating it with other OWL ontologies, such as the NCI thesaurus [153]. The problem is that frames' semantics is not as precisely defined as Description Logics' one. Moreover, although superficially similar, these two approaches rely on fundamentally different modeling assumptions, and there is no direct mapping between them. Protégé[4], the ontology editing platform that was used to build the FMA supports both formalisms. The frame-based mode has an "export to OWL" option. However, this option only performs a straightforward translation

[4] `http://protege.stanford.edu/`

that ignores all the features that do not have a direct equivalent. Moreover, it does not take advantage of all the OWL-specific features that are the basis of the language strength. For these two reasons, the resulting translation would not be usable for reasoning.

3.2.2 Objective

We analyze some theoretical and computational issues of representing the FMA in OWL-DL. To address the expressiveness limitation, we propose to use a more expressive formalism ensuring application independence while meeting the expressiveness requirements. To address the computational limitations, we propose a "Virtual FMA-OWL" architecture based on a Web Service that returns the OWL-Full representation of a concept given its identifier. Eventually, we advocate the use of this architecture for continuing to maintain the FMA in the current frame-based form while making it accessible to the Semantic Web. Note that the intention of this article is not to discuss the modeling of the FMA [151], but rather to examine different representation formalisms considering the computational requirements of the applications that use them.

3.2.3 Converting the FMA into OWL-DL

The FMA is currently composed of more than 70,000 anatomical items called concepts, having more than 1.5 million relationships (such as composition, neighborhood or blood supply) between them. The concepts are identified by a unique number called the FMAID, and are associated with one or more designation (e.g. the string "Heart" for the concept 7088 corresponding to the heart), which allows to handle synonyms or multiple languages. The concepts are strictly organized in a principled specialization hierarchy [152].

3.2.3.1 Basic concept representation: identifiers and designations

We represented the FMA concepts as OWL classes, and relationships as OWL properties. Classes were identified by their FMAID, relative to the FMA namespace[5]. This allows us to avoid any potential ambiguity with another ontology having a concept with the same identifier, as different ontologies have different namespaces. We used RDF labels to represent the concept designation, explicitly mentioning the language. This is illustrated by the Figure 3.2, in which the identifier is interpreted against the default namespace, which is declared at the ontology level to be the FMA one).

```
<owl:Class rdf:ID="7088">
  <rdfs:label xml:lang="en">Heart</rdfs:label>
  <rdfs:label xml:lang="fr">Coeur</rfds:label>
  ...
</owl:Class>
```

Figure 3.2: Representation of the FMA identifiers and designations in OWL-DL.

3.2.3.2 Taxonomy and metaclasses

The FMA features a complex structure of superclasses and subclasses [152]. For example, "Physical anatomical entity" is an instance of "Anatomical entity template", and a subclass of both "Anatomical entity template" and "Anatomical entity".

[5] http://sig.biostr.washington.edu/fma#

The representation of the original FMA taxonomy in OWL was straightforward. The subclasses—superclass relation between frames was represented by the *rdfs:subClassOf* relation. The resulting hierarchy is homologous to the original FMA one (see Figure 3.3).

OWL-DL does not support metaclasses, so we needed to remove them.

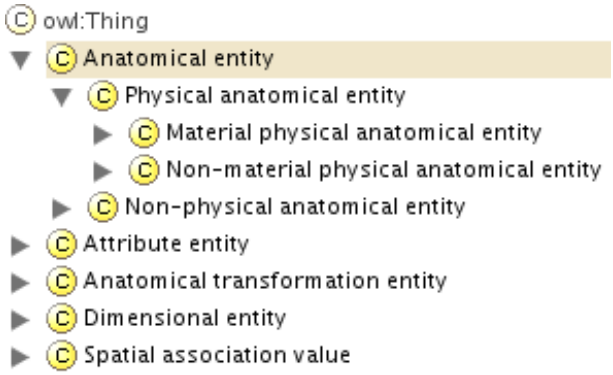

Figure 3.3: Taxonomy of the OWL-DL representation of the FMA.

3.2.3.3 Disjointness

Description Logics' modeling principles are slightly different from those of frames. These differences have to be taken into account during conversion. Particularly, the FMA is organized in a hierarchy of mutually-disjoint concepts. However, in Description Logics (hence in OWL-DL), classes are not disjoint by default (i.e. there can exist an individual that is an instance of both classes). Therefore, in order to respect the FMA modeling principles, we assume that unless specified otherwise by multiple inheritance, all the direct subclasses of a class are mutually disjoint. For example, Esophagus and Stomach are two direct subclasses of "Organ with organ cavity" and they are disjoint (an instance of "Esophagus" cannot be also an instance of "Stomach"). However, "Left breast", "Right breast", "Male breast" and "Female breast" should not be specified as disjoint (although they are automatically because "Left female breast" is only described as a subclass of "Female breast", and not also of "Left breast"). This point will be further discussed in section 3.2.5.

Note that this knowledge was implicit in the frames version of the FMA and is made explicit in its OWL version.

3.2.3.4 Closure

Another difference between frames and Description Logics is that the latter relies on the "open world assumption" [146] whereas the former assumes a closed world. In a closed world, everything that is not explicitly stated is assumed to be false.

Consequently, when the FMA describes the parts of an anatomical structure such as the hand, the fact that the structures other than the palm or the finger are not said to be parts of the hand is interpreted as "they are not part of the hand". However, in Description Logics, providing a list of the possible parts of the hand does not prevent other structures to be also parts of the hand. Therefore, we have to add an extra constraint saying that the structures in the list are the only possible parts of the hand. This is called introducing a closure axiom [154], and it has to be done for all the relationships (for an example see [73]).

However, generating closures is much more complicated than it may seem at first sight. For example, the possible parts of the hand are the palm and the five fingers. Now, we have to

take overloading by subclasses into account so that the possible parts of the "Left hand" are "Left palm", "Left thumb", ..., "Left little finger" (note that the closure does not mention non-lateralized concepts such as "Thumb" anymore). However, the same approach cannot be applied to the lungs: a lung has an upper lobe and a lower lobe as parts. Its subclass "Right lung" not only has parts "Upper lobe of right lung" and "Lower lobe of right lung" (same approach as for the hand), but also a middle lobe that does not exist for the left lung. As a consequence, it would be incorrect to generate a closure for the lung based on its parts, whereas it should be done for the hand. The first situation occurs when the child overloads its parent. The second one occurs when subclasses introduce new properties. Unfortunately, real world situations can mix these two situations.

In order to automate the systematic generation of closures, we have to check if all the classes that define the range of a relation for a concept are subclasses of the range of this relation for the superclasses of the concept. This point will be further discussed in section 3.2.5.

3.2.3.5 N-ary and attributed relationships

N-ary relationships associate more than two entities. Particularly, this is extensively used in the FMA to qualify a relation between two entities. Those attributed relationships are used to qualify part or continuity relationships for example (the lung is continuous medially to the pulmonary veinous tree).

The modeling in Description Logics of such relationships has been studied by the W3C Semantic Web Best Practice working group, and we followed their recommendation [155].

3.2.4 Addressing expressiveness and application-independence: OWL-Full

From the previous section, we have seen that some of the FMA features are simply out of the scope of OWL-DL. We propose a two-layered approach. The first layer consists of a generic conversion tool that generates a representation of the FMA in OWL-Full. The second layer consists of several application-specific optimization tools that simplify the OWL-Full representation of concepts into OWL-DL ones by removing all the features unnecessary according to the application context.

OWL-Full does not suffer from the expressiveness limitations of OWL-DL. For example, it supports metaclasses. Figure 3.4 shows that the "Heart" (FMAID: 7088) can be represented in OWL-Full both as a subclass and as an instance of "Organ with cavitated organ parts" (FMAID: 55673), which complies with the original FMA structure.

```
<owl:Class rdf:ID="7088">
  <rdfs:label xml:lang="en">Heart</rdfs:label>
  <rdfs:label xml:lang="fr">Coeur</rfds:label>
  <rdfs:subClassOf rdf:resource=
     "http://sig.biostr.washington.edu/fma#55673"/>
  <rdfs:type rdf:resource=
     "http://sig.biostr.washington.edu/fma#55673"/>
  ...
</owl:Class>
```

Figure 3.4: Representing the original FMA metaclasses and subclasses structure in OWL-Full (concept `7088` is the heart; concept `55673` is "Organ with cavitated organ part").

OWL-Full allows us to generate a layer that has all the expressiveness we may need and that

is application-independent. Moreover, this approach promotes interoperability: any application that requests the concept 7088 gets the same description in OWL-Full. The application is then free to modify this description internally according to its specific needs (namely simplify it to meet its computational requirements), but at least, the communication between applications refers to a shared representation.

3.2.5 Pattern-based generation of consistency constraints

This section summarizes a collaboration with Julie Chabalier between 2006 and 2007, and is therefore posterior to the initial work from 2004. Previous works showed that some features are implicit or cannot be represented in frames but are crucial for leveraging the specificities of OWL [75, 156, 157, 158]. It would be possible to address this point by manual processing, but the task is likely to be cumbersome, error-prone, and would increase the workload of maintaining the original FMA. Moreover, the organization of the FMA follows a strict and principled approach that could be exploited.

We identified in the original FMA a set of patterns reflecting situations with an underlying modeling principle that could not be partially or totally represented in frames, and is therefore missing. We wrote a set of Python scripts for detecting these patterns among the classes of the original FMA and generating the corresponding OWL constraints.

3.2.5.1 Representing multiple inheritance

The FMA taxonomy follows a very principled approach. We duplicated this taxonomy in OWL (cf. section 3.2.3.2). However, because the original FMA only uses single inheritance and because the distinction between single and multiple inheritance is not relevant in OWL, we generated some additional taxonomic relationships.

Due to the single inheritance constraint, the following situations incompletely account for the subclass—superclass structure:

- left/right and male/female: 3 classes (`Breast`, `Areola`, `Nipple`)
- left/right and enumeration: 65 classes (e.g. `Left first cervical nerve`)
- upper/lower and enumeration: 5 classes (e.g. `Upper first molar socket`)

For example, `Breast` has four direct subclasses: `Left breast`, `Right breast`, `Male breast` and `Female breast`. Each of `Male breast` and `Female breast` have one left and one right direct subclass. Consequently, `Left male breast` is a subclass of `Male breast`, but not of `Left breast`. Moreover, classes such as `Intercostal lymph node` combine the last two patterns.

3.2.5.2 Disjointness

In the original FMA, by default, all the sibling classes are disjoint. For example, the direct subclasses of `Cell` are `Nucleated cell` and `Non-nucleated cell`, and it is clear that a cell cannot be at the same time nucleated and non-nucleated. This feature can be made explicit in OWL with disjointness constraints.

However, systematically making all the sibling classes mutually disjoints requires to rule out situations where all the direct subclasses of a class are not mutually exclusive. In the previous example, `Left breast` and `Right breast`, as well as `Male breast` and `Female breast` are disjoint, but `Left breast` and `Male breast` are not. Similarly, the class `Region of chest wall` has seven direct subclasses, including `Anterior chest wall`, `Superficial chest wall`, `Anterior superficial chest wall`, `Lateral chest wall` and `Lateral superficial chest`

wall. The problem was then to distinguish among the siblings the pairs of disjoint classes from those that are not.

Rather than risking some inconsistency or some trivial satisfiability of the ontology, we chose the conservative approach of only generating the disjointness constraints we are certain of. This lead us to the identification of some other patterns among sibling classes:

- `Left X/Right X`: 3736 classes (e.g. `Left lung`)
- `X left Y/X right Y`: 13989 classes (e.g. `Skin of right breast`)
- `Male X/Female X`: 25 classes (e.g. `Male breast`)
- `X male Y/X female Y`: 75 classes (e.g. `Right side of male chest`)
- enumeration: XX classes (e.g. `First cervical nerve`)
- upper/(middle)/lower: YY classes (e.g. `Upper lobe of lung`)

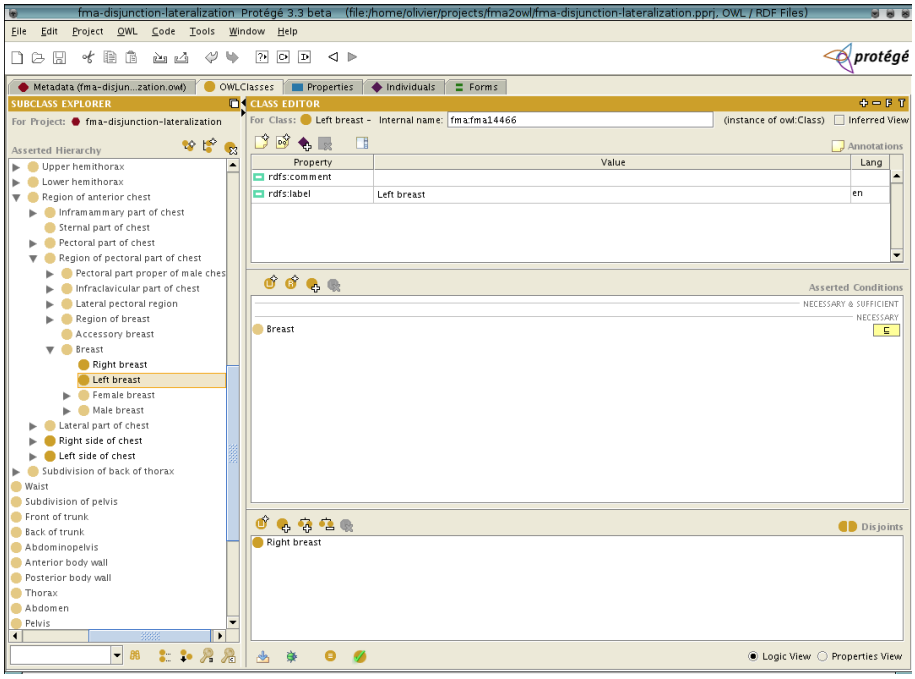

Figure 3.5: Disjointness axiom for `LeftBreast` stating that it is disjoint from `RightBreast`. It was generated with the `Left X/Right X` pattern.

3.2.5.3 Necessary and sufficient definition

Coverage The FMA aims at completeness. It is assumed that for each class, its subclasses provide a complete decomposition (i.e. there are no `X-Other` nor `X-Unspecified` subclasses of the `X` class). For example, the two direct subclasses of `Organ` are `Solid organ` and `Cavitated organ`. The intended meaning is that each organ is either solid or cavitated, and that there is no third possibility. In OWL, this can be made explicit by a coverage definition. In the previous example, `Organ` would be defined as the union of `Solid organ` and `Cavitated organ`.

A naive approach would consist in generating a coverage definition for each class using the union of its direct subclasses. This would successfully generate the definition that a lobe of lung is either an upper lobe of lung or a middle lobe of lung or a lower lobe of lung. The result would always be correct from a logical point of view. However, in some situations, it can still be refined. If we return to the class `Breast`, we could further specify that a breast is either a left breast or a right one, and also that it is either a male breast or a female one.

In order to generate coverage definitions as specific as possible, we reused the patterns identified for generating multiple inheritance (Section 3.2.5.1) and disjointness (Section 3.2.5.2). For each of the matching pattern, we generated the corresponding coverage (this is what allows us to generate two definitions for `Breast`, cf. Figure 3.6). We also generated a coverage definition with the classes matching none of the patterns (e.g. for `Lobe of lung`).

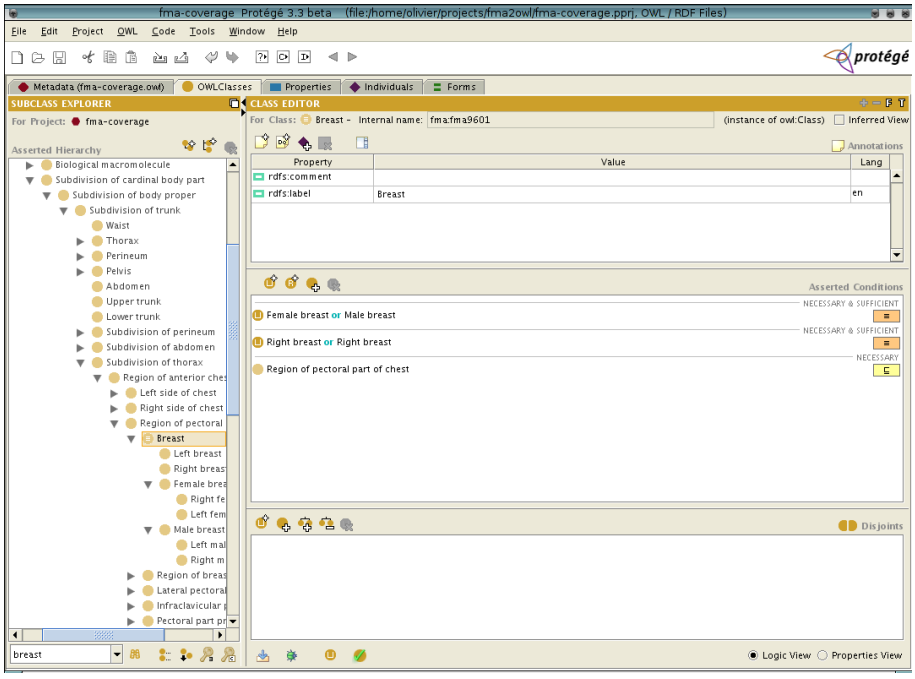

Figure 3.6: Two coverage axioms for `Breast` represented as necessary and sufficient definitions for `Breast`. They were generated with the `Left X/Right X` and the `Male X/Female X` patterns.

3.3 Methodology: diagnosis of heart-related injuries

This study aims at improving diagnosis and prognosis of battlefield injuries in the region of the heart. It focuses on a reasoning method based on the FMA for (1) determining which parts of the heart muscle will become partially or totally ischemic in case of an injury involving a coronary artery, and (2) determining whether a perforation of the wall of the heart will lead to massive bleeding or will be (temporarily) contained by the surrounding cavity. Both scenarios perform some semantically-rich reasoning. The first one involves class-based reasoning, and the second one involves instance-based reasoning. Together, they cover most of OWL1.0 constructs.
In retrospect, this work is interesting because it demonstrated that semantically-rich ontologies represented in Description Logics actually support advanced reasoning. It also showed that once the ontologies are available (which is a critical limitation, we have covered some of the related difficulties in section 3.2), developing the application-specific part does not require a large amount of work (it was a matter of hours). Eventually, the encapsulation of the symbolic reasoning into a Web service was relevant in terms of software architecture and showed that the end user does not have to operate an ontology editor.

This work was a contribution to DARPA's Virtual Solder project[6] during my postdoc with Mark Musen. It was originally published in: Daniel L. Rubin, Olivier Dameron, and Mark A. Musen. Use of Description Logic classification to reason about consequences of penetrating injuries. In *American Medical Informatics Association Conference AMIA05*, pages 649–653, 2005 [73]. It was also a contribution to: Daniel L. Rubin, Olivier Dameron, Yasser Bashir, David Grossman, Parvati Dev, and Mark A. Musen. Using ontologies linked with geometric models to reason about penetrating injuries. *Artificial Intelligence in Medicine*, 37(3):167–176, 2006 [74].

3.3.1 Context

The Virtual Soldier project developed complex mathematical models to create physiological representations of individual soldiers that can be used to improve medical diagnosis on and off the battlefield. Soldiers would be equipped with "P tags" that are USB-like storage devices containing their 3D anatomical information, as well as physiological and biological parameters, their genetic information and their medical record. They would also be typically equipped with "intelligent" battledress and sensors that monitor their vital signs and physiological parameters and would also be able to record the location of a bullet entry and exit points as well as some associated parameters such as bullet velocity. All these informations can be collected by medical teams in the field and provided to some remote diagnosis and prognosis decision support system.

Primary penetrating injuries concern the anatomical structures directly impaired by the internal trajectory of a bullet. They can be determined with spatial geometric models of injured subjects by computing the intersection of the bullet's cone of damage with the organs. *Secondary injuries* are consequences of primary injuries. Typically, if the cone of damage impairs an artery, the organs perfused by the artery will experience ischemia. The determination of secondary injuries relies on background knowledge about anatomy (e.g. the Foundational Model of Anatomy [151]) and physiology (e.g. The Foundational Model of Physiology [159]).

A challenge in creating new decision support systems is to incorporate medical knowledge and to apply that knowledge in flexible ways [160]. In most reasoning systems that use ontologies, the

[6] `http://www.virtualsoldier.us/`

knowledge used to guide reasoning (control knowledge) is embedded in the application code or in rules used in conjunction with the domain ontology [161]. We believe that it is advantageous to use Description Logics in biomedical applications to represent both the domain knowledge and the control knowledge needed for reasoning. Thus, we construe the reasoning problems in the domain as classification tasks.

3.3.2 Objective

Given a set of anatomic structures that are directly injured by a projectile, we want to create a reasoning application that deduces secondary injuries of two types: (1) regions of myocardium that will be ischemic if a coronary artery is injured, and (2) propagation of injury as bleeding occurs into damaged anatomic compartments that surround the heart.

We modeled these tasks as classification problems. We describe our approach to creating reasoning services that fulfill the above desiderata using OWL. In this work we exploit the automated reasoning capability provided by OWL.

3.3.3 Reasoning about coronary artery ischemia

We created a reasoning service to infer the myocardial ischemic consequences of coronary artery injury ("Cardiac Ischemia Reasoner"). This service relied on class-based reasoning, i.e. we modeled the query as a set of defined classes, and we checked which anatomical entities were inferred to be subclasses of these defined classes.

3.3.3.1 Modeling blood vessels

We added necessary and sufficient conditions to classes in our base OWL ontology of anatomy to encode the dependency of downstream arterial branches on the upstream arteries, and to represent the regions of the heart myocardium supplied by the coronary artery branches (Figure 3.7 on the following page). For example, we represented the composition of the coronary arteries using the *hasSegment relation*, and the tree-like structure of the blood vessels with the *isContinuousWithOutputOf*.

We created a new primitive class `SeveredBloodVessel` as a subclass of `BloodVessel`. This class will serve as an input for the reasoning service: when the geometric analysis detects that a primary injury involves a blood vessel, we declare that the corresponding class is a subclass of `SeveredBloodVessel` (therefore it remains an indirect subclass of `BloodVessel`).

$$\texttt{SeveredBloodVessel} \sqsubset \texttt{BloodVessel}$$

We created a new defined class `FunctionallyImpairedBloodVessel` to infer that all the blood vessels downstream a severed blood vessels are also affected. Figure 3.8 on the next page shows that after declaring that the second segment of the right coronary artery was injured, all its downstream branches are inferred to be functionally impaired.

$$\texttt{FunctionallyImpairedBloodVessel} \equiv \begin{cases} \texttt{SeveredBloodVessel} \\ \sqcup \\ (\exists \textit{isContinuousWithOutputOf} \\ \texttt{FunctionallyImpairedBloodVessel}) \end{cases}$$

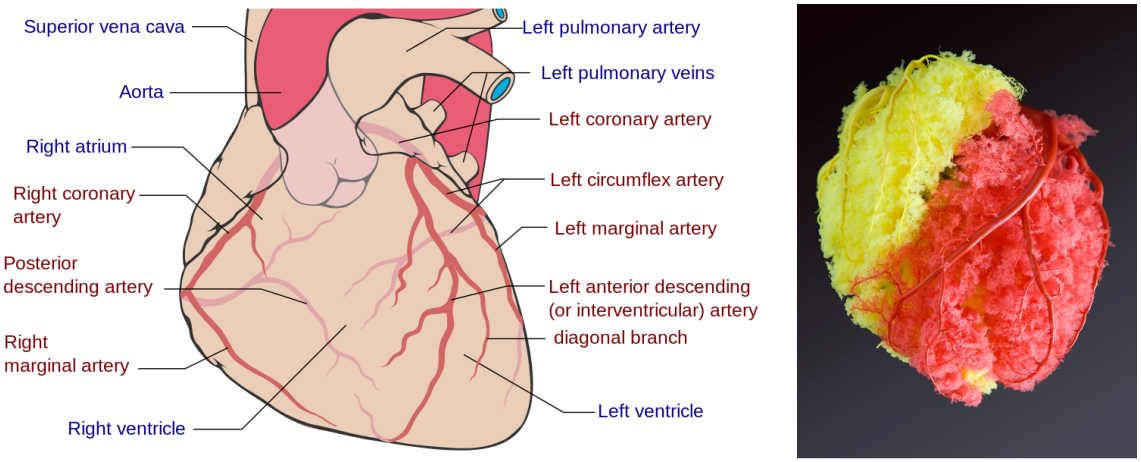

Figure 3.7: Schema of the coronary arteries and their branches, and cast of the coronary arteries (yellow: right coronary artery; red = left coronary artery) showing the regions of the myocardium they provide blood to. Left schema is from `http://en.wikipedia.org/wiki/File:Coronary_arteries.svg` under the CC-BY-SA license. Right image is from `http://en.wikipedia.org/wiki/File:Coronary_Arteries.tif` and is in the public domain.

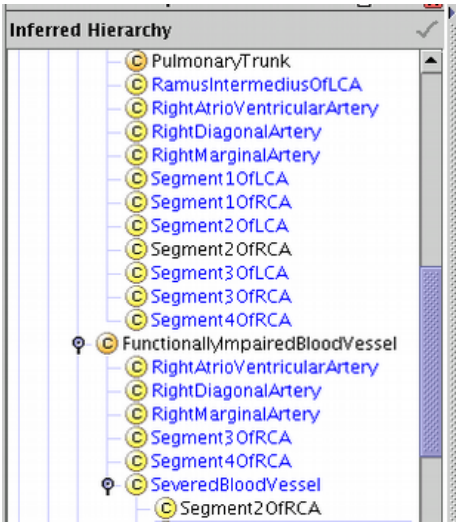

Figure 3.8: Inference that all the downstream branches of the second segment of the right coronary artery are functionally impaired after it has been injured.

3.3.3.2 Describing blood supply to organs

To represent the coronary arteries that supply the lateral part of the wall of the left ventricle, we added restrictions to the class `LateralPartOfWallOfLeftVentricle` that specify values for the *isSuppliedBy* property, such as `LeftCircumflexArtery` (Figure 3.9). Note the closure axiom indicating that the left circumflex artery, the ramus intermedius and the diagonal branch of the left coronary artery are the only blood vessels supplying the lateral part of the wall of the left ventricle.

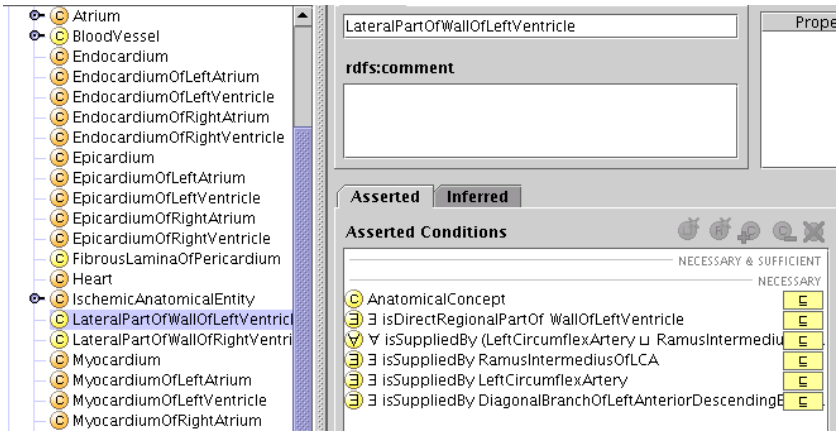

Figure 3.9: OWL Ontology of coronary anatomy and regional myocardial perfusion. Classes of anatomic structures are shown in the left panel, and logical definitions of the concepts are on the right. The class `LateralPartOfWallOfLeftVentricle` contains six restrictions representing the necessary conditions for this class. Some of these assertions specify the coronary arterial branches that supply this structure.

3.3.3.3 Modeling ischemia

We created a defined class `IschemicAnatomicalEntity` as an `AnatomicalEntity` that is supplied by at least one functionally impaired blood vessel.

$$\texttt{IschemicAnatomicalEntity} \equiv \begin{cases} \texttt{AnatomicalEntity} \\ \sqcap \\ (\exists \textit{isSuppliedBy}\, \texttt{FunctionallyImpairedBloodVessel}) \end{cases}$$

An organ may be supplied by more than one artery, in which case damage to one of the feeding arteries will cause partial (not complete) impairment of blood flow to the organ. To represent these types of ischemia, we refined `IschemicAnatomicalEntity` into two defined subclasses `IschemicAnatomicalEntityPartially` and `IschemicAnatomicalEntityTotally`. Figure 3.10 on the next page shows the inferred ischemic anatomical entities after the second segment of the right coronary artery has been severed. The posterior wall of the left ventricle is partially ischemic because it is also supplied by the left coronary artery. Note that two anatomical entities (right atrium and posterior wall of the right ventricle) are correctly inferred to be ischemic but the system could not determine whether they were partially or totally ischemic. For these two anatomical entities, we had omitted to specify a closure axion (cf. Fig 3.9) on purpose to demonstrate why closure is important when using open-world reasoning (more on this in chapter 4)

$$\texttt{IschemicAnatomicalEntityPartially} \equiv \left\{ \begin{array}{l} \texttt{IschemicAnatomicalEntity} \\ \sqcap \\ (\exists \textit{isSuppliedBy}\, \texttt{FunctionallyNonImpairedBloodVessel}) \end{array} \right.$$

$$\texttt{IschemicAnatomicalEntityTotally} \equiv \left\{ \begin{array}{l} \texttt{IschemicAnatomicalEntity} \\ \sqcap \\ (\forall \textit{isSuppliedBy}\, \texttt{FunctionallyImpairedBloodVessel}) \end{array} \right.$$

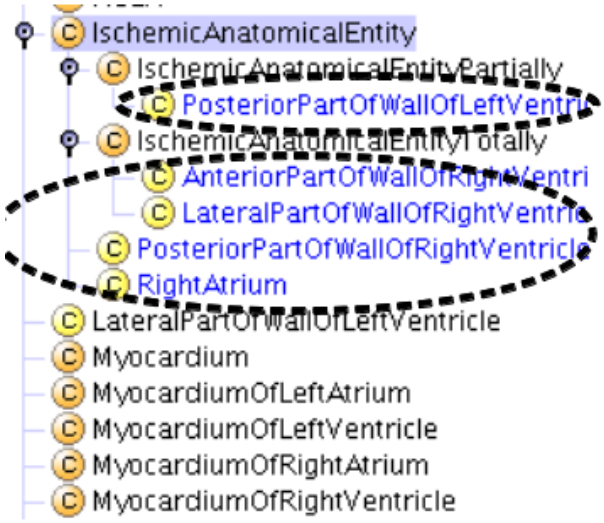

Figure 3.10: Cardiac Ischemia OWL ontology updated with the knowledge that the second segment of the right coronary artery has been injured. After automatic classification, particular anatomic classes (circled) are reclassified, suggesting the ischemic regions of myocardium that occur as a consequence of the right coronary artery injury.

3.3.4 Reasoning about pericardial effusion

We created a second reasoning service to infer the cavities affected by bleeding after an injury ("Injury Propagation Reasoner"). The heart is surrounded by two membranes – the pericardium and the pleura, that determine two cavities enclosed inside each other (the heart is enclosed in the pericardial cavity, which is enclosed in the pleural cavity) and are normally filled with serous fluid. The (abnormal) presence of blood in these cavities is known as hemopericardium and hemothorax (Figure 3.11 on the next page). In certain cases, the increased pressure can limit hemorrhage.

This service relied on instance-based reasoning, i.e. we modeled the patient's condition by creating instances of all the relevant anatomical structures and by linking these instances with the appropriate relations. We modeled the query as a set of defined classes, and we checked which anatomical entities were inferred to be instances of these defined classes.

We first indicated in the ontology that blood vessels and cardiac cavities (the left and right atrium and ventricles) are filled with blood (Figure 3.12 on the facing page).

In order to represent a perforation in the wall of the heart, we created an instance of the class `AddedConduit`, and added values to the *continuousWith* property to describe that this conduit

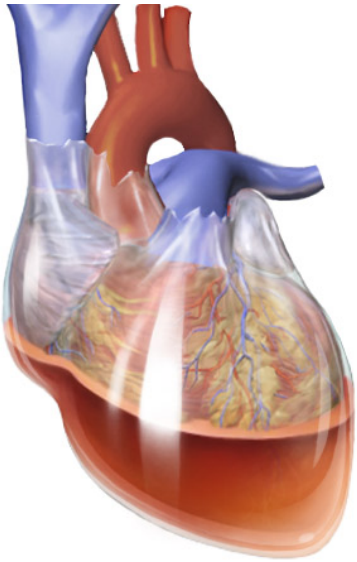

Figure 3.11: Blood loss through a punctured cardiac membrane fills the pericardial cavity. This can lead to a massive hemorrhage or on the contrary provide some temporarily containment. Schema from `http://en.wikipedia.org/wiki/File:Blausen_0164_CardiacTamponade_02.png` under CC-BY license.

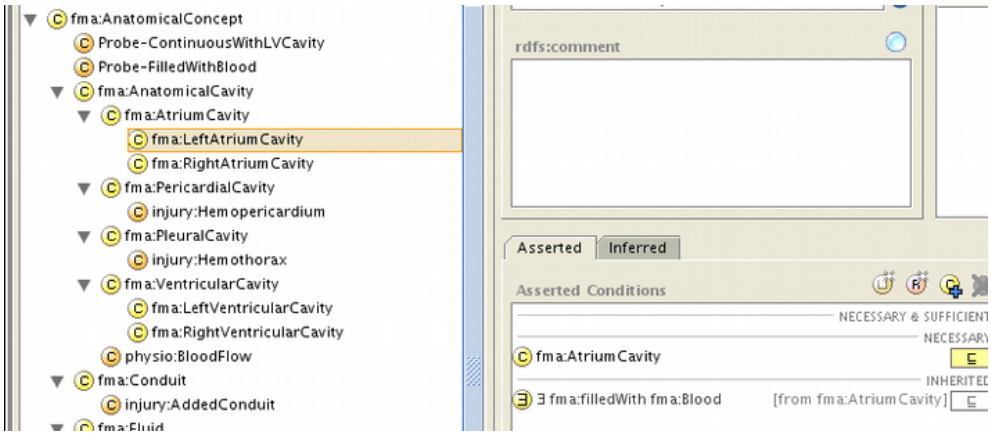

Figure 3.12: Blood vessels and cardiac cavities are filled with blood (and this is normal).

connects the cavity of the left ventricle and the pericardial space (Figure 3.13). The *continuousWith* property represents spatial continuity between adjacent hollow anatomic structures that have been injured. It is symmetric and transitive. These two property characteristics are needed to infer that, given a perforation in the wall of the left ventricle (`HoleInWallOfHeart`) and pericardium (`HoleInPericardium`) creating conduits that connect the surrounding cavities, the conduits, pericardial cavity, and pleural cavity will be in continuity with the cavity of the left ventricle (Figure 3.14).

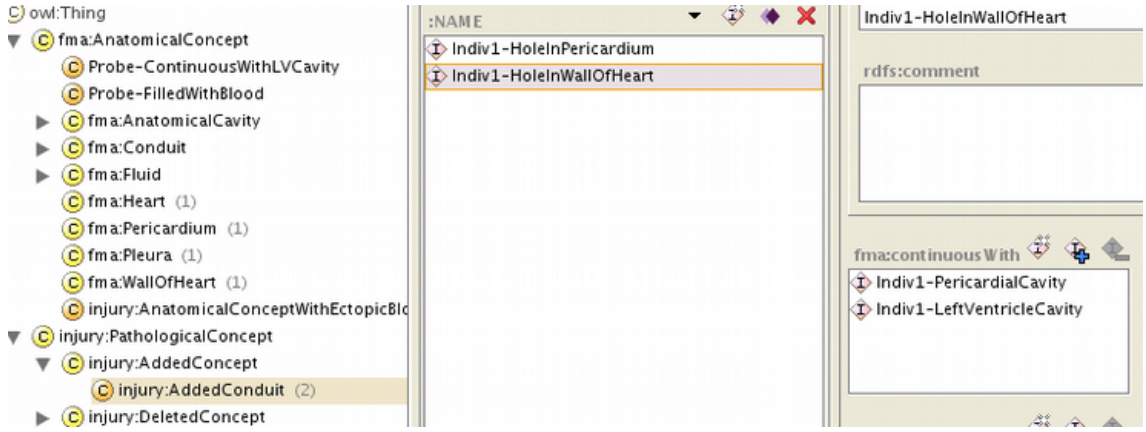

Figure 3.13: Knowledge representation in OWL of a hole in the heart wall. An instance of the `AddedConduit` class is created, having values of the *continuousWith* property specifying the anatomic compartments connected by this conduit.

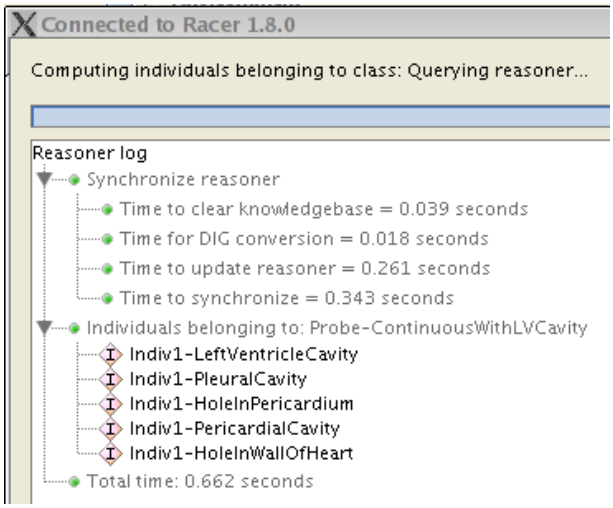

Figure 3.14: Inferred knowledge after asserting a cardiac injury comprising a hole in the left ventricle and classifying the Injury Propagation OWL ontology. The pericardial cavity and pleural cavity are inferred to be in continuity with the left ventricle.

We created a defined class `BloodFlow` (Figure 3.15 on the next page). The necessary and sufficient condition of the `BloodFlow` class defines that any anatomical cavity continuous with something filled with blood is an instance of `BloodFlow`. The necessary condition indicates that if an individual is an instance of `BloodFlow`, then it is itself filled with blood. The combination of these two conditions models the rule "if a cavity is continuous with something filled with blood, then it is itself filled with blood". In order to check that the reasoning was correct, we

created a defined probe class to retrieve the anatomical entities filled with blood (Figure 3.16). For the cardiac cavities, being filled with blood is a good thing, whereas for the pericardial and the pleural cavities, this is abnormal. In order to detect cavities abnormally filled with blood, we defined the class `Hemopericardium` as a pericardial cavity that happens to be filled with blood, and we make it a subclass of the things abnormally filled with blood (Figure 3.17 on the next page). We repeated the process with the pleural cavity and created a `Hemothorax` class. Note that we used the same rule pattern as with `BloodFlow`. Eventually, we created the class `AnatomicalConceptWithEctopicBlood` to retrieve the places abnormally filled with blood (Figure 3.18 on the following page).

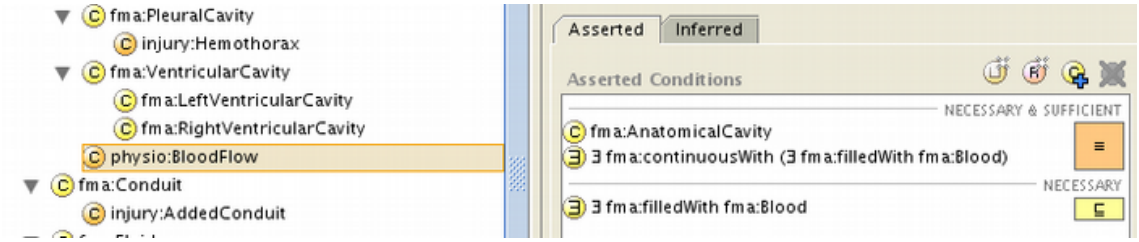

Figure 3.15: The `BloodFlow` class uses a necessary and sufficient definition and a necessary condition to model the rule "if a cavity is continuous with something filled with blood, then it is itself filled with blood".

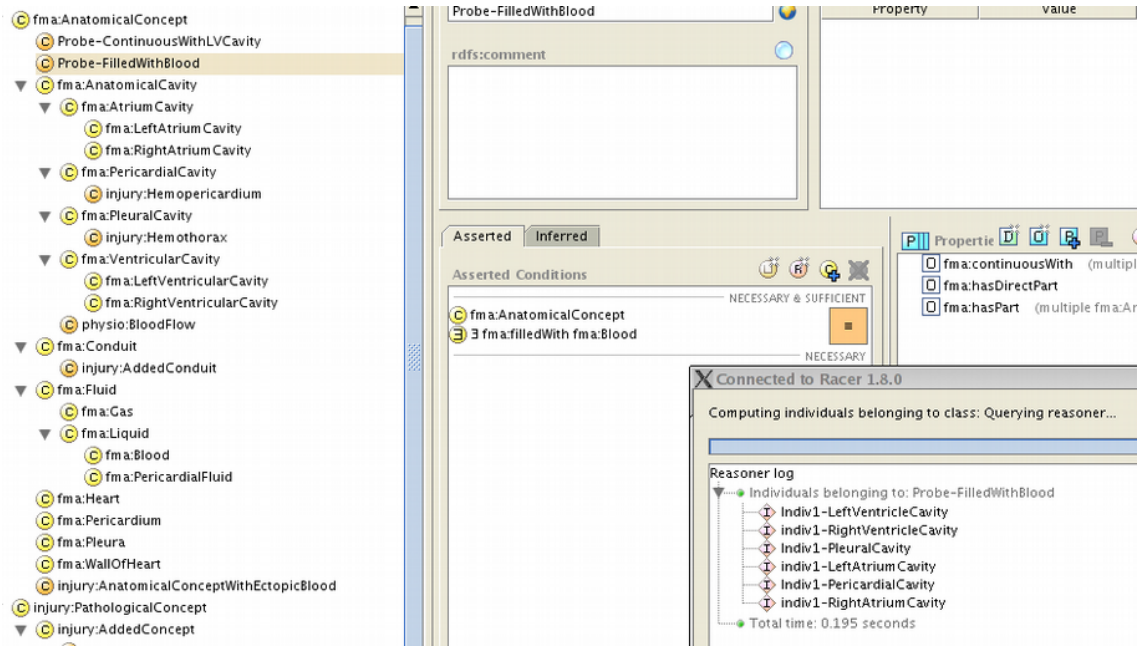

Figure 3.16: A probe class indicates the anatomical entities filled with blood. Note that the system inferred correctly that after the injury the pericardial and the pleural cavities are filled with blood.

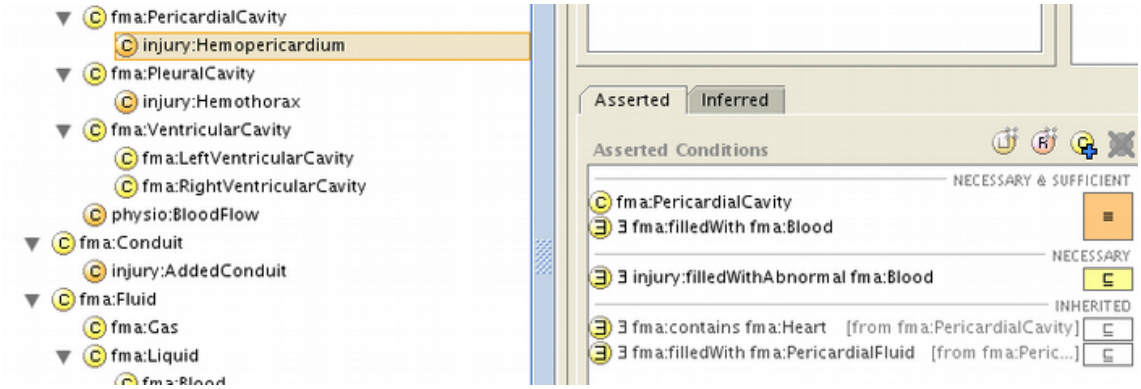

Figure 3.17: `Hemopericardium` is defined as a pericardial cavity filled with blood. The necessary condition ensures that any instance of this class is then inferred to be abnormally filled with blood.

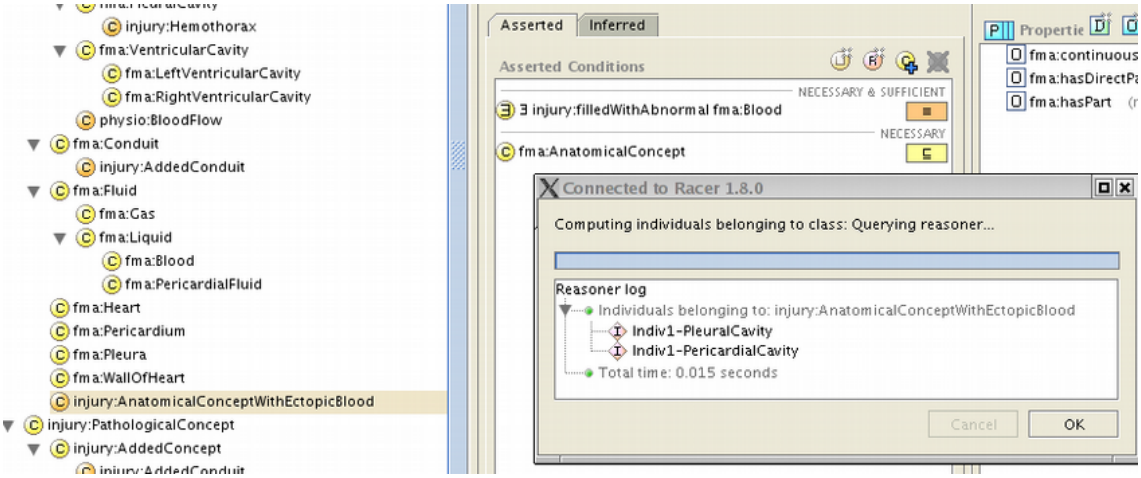

Figure 3.18: The `AnatomicalConceptWithEctopicBlood` retrieves the anatomical entities abnormally filled with blood.

3.4 Optimization: modeling strategies for estimating pacemaker alerts severity

This study focuses on the determination of the best modeling strategy (in terms of correctness and performances) for predicting the severity level of a pacemaker alert. Contrary to the previous studies, this application potentially involved processing a large number of alerts, so performances became important. The previous section with the Virtual Soldier project already demonstrated that some problems can be modeled with class-based or instance-based OWL reasoning, but provided no hint on whether one of them is better than the other. It is not even clear that statements such as "class-based reasoning is faster than instance-based reasoning" actually make any sense, as the outcome may depend on the problem at hand. This empirical study went one step further by considering SWRL rules in addition to OWL, and by exploring systematically all the combinations of modeling strategies. The results showed that both OWL and SWRL-based ontology modeling techniques can reliably perform the reasoning necessary to propose a severity level associated with pacemaker alerts. The best performances were not obtained by using exclusively OWL nor SWRL but by combining their respective advantages, using OWL to reduce the number of SWRL rules and making them simpler.
In retrospect, this work is interesting because it was the first hands-on systematic evaluation of modeling strategies. While applications usually use exclusively SPARQL, OWL or SWRL, optimizing performances suggests to combine their various strength. By extrapolation, this suggests a more formal approach for designing complex reasoning tasks as workflows instead of monolithic entities, which allows to choose the best technology for each module and to promote their reuse.

This study was a contribution to the Akenaton[7] project (ANR-07-TECS-0001). It was originally published in: Olivier Dameron, Pascal van Hille, Lynda Temal, Arnaud Rosier, Louise Deléger, Cyril Grouin, Pierre Zweigenbaum, and Anita Burgun. Comparison of OWL and SWRL-based ontology modeling strategies for the determination of pacemaker alerts severity. In *Proceedings of the American Medical Informatics Association Conference AMIA*, page 284, 2011 [92], where it was shortlisted for the best article award.

It was later extended in: Pascal van Hille, Julie Jacques, Julien Taillard, Arnaud Rosier, David Delerue, Anita Burgun, and Olivier Dameron. Comparing Drools and ontology-based reasoning approaches for telecardiology decision support. *Studies in health technology and informatics*, 180:300–304, 2012 [93] where we also considered Drools rules. Drools rules are not capable of handling the granularity gap between precise patients data and more general criteria, so I do not present this later work here. We had to generate Drools rules that mimicked ontology-based reasoning. The conclusion was that the limitations of ontology-based reasoning were the reasoner's performances, whereas the limitations of Drools were the number and complexity of rules. This suggested using ontology for automatically generating some of the Drools rules.

3.4.1 Context

Patients suffering from heart failure are increasingly treated with implantable cardioverter defibrillators (ICD) and benefit from home monitoring [162]. In this context of telecardiology, ICDs send remote alerts about arrhythmic episodes to physicians, who have to determine their

[7] `http://www.agence-nationale-recherche.fr/?Project=ANR-07-TECS-0001`

emergency level and potentially take the required actions [163, 164]. Some automatic triage of the alerts according to their emergency level is instrumental to keep up with this overwhelming flow of alerts (from zero most of the time up to as many as twenty alerts per patient per day; with an estimation of 500.000 new patients every year) efficiently. However, this is an intrinsically difficult task because the risk associated with an alert depends on multiple interdependent factors such as the patient's medical history, his current pathologies and his current treatment [165]. For example, in case of atrial fibrillation (AF), the risk of thrombo-embolism is estimated by the CHA2DS2VASc score as well as by additional parameters that can either increase the risk (e.g., if the patient is a smoker or is obese) or lower it (e.g., if the patient is currently treated with an anticoagulant) [166, 167].

The goal of the AKENATON project is to improve ICD alert management by automatically associating each alert with a severity level [91]. This requires (i) to extract the relevant data from the alerts transmitted by ICDs and from the patient's clinical context, (ii) to integrate them, (iii) to reconcile them with the severity criteria, and (iv) to compute the alert severity. Extracting data relies on queries on the hospital patient database as well as on Natural Language Processing techniques for mining free text and structured documents. Integrating data and reconciling the granularity gap with more general severity criteria requires symbolic domain knowledge represented as ontologies (e.g., in order to automatically recognize that a patient suffering from a right iliac artery stenosis will match the vascular disease CHA2DS2VASc criterion). Deducing the alert severity from the various criteria is a combination of ontology-based and rule-based inferences. The ontology model plays a central role in several steps and ensures the general coherence. It can be represented using different combinations of OWL [168] and SWRL [169]. Each combination implies specific modeling decisions which may have consequences on the performance of the system. Currently, no guideline exists to decide which strategy best fits our particular problem.

3.4.2 Objective

This study focuses on the determination of the best ontology modeling strategy to integrate data and to fill the granularity gap between data and the CHA2DS2VASc score criteria for patients with an atrial fibrillation alert.

First, we identified the CHA2DS2VASc criteria that potentially require reference to domain knowledge to be reconciled with patient data. Second, we identified ten modeling strategies covering all the possible combinations of Java, OWL-DL and SWRL. For each strategy, we assessed the modeling effort by counting the number of OWL classes, properties and SWRL rules. Third, we validated each strategy by verifying that they computed the correct score for all of the 192 possible combinations of criteria. Fourth, we compared the performances of the ten strategies by measuring the computation time for each 192 cases of the validation set. Fifth, we evaluated all the strategies over a corpus of 62 actual patients by repeating steps three and four.

3.4.3 CHA2DS2VASc score

CHA2DS2VASc is a new recommendation of the European Society of Cardiology to determine stroke risk for patients with non-valvular fibrillation [166, 167, 170]. The higher the CHA2DS2VASc score, the higher the risk of thrombo-embolism [166, 167]. It is a major determinant for deciding whether or not an anticoagulation therapy is required in order to prevent potential stroke caused by stasis of blood in the heart, which may lead to the formation of a thrombus that can dislodge into the blood flow. A CHA2DS2VASc score of zero is associated with a low risk, a score of one is associated with an intermediate risk, and a score of two or

more is associated with a high risk [167]. Table 3.1 presents the criteria required to compute the CHA2DS2VASc score. Some of them such as age or sex category can be directly computed from the patient's administrative data. Medical criteria such as congestive heart failure or vascular disease are more general. They encompass several diseases and are therefore unlikely to be present as such in the patient's data. Reconciling the patient's data with the CHA2DS2VASc criteria consists in interpreting the data according to some domain-specific knowledge, typically represented in ontologies.

Criterion	Points
Congestive heart failure / left ventricular dysfunction	1
Hypertension	1
Age $\geq$ 75 y.o.	2
Diabetes mellitus	1
Stroke / transient ischemic attack / thromboembolism	2
Vascular disease	1
65 $\leq$ **A**ge $<$ 75	1
Sex **c**ategory (ie, female gender)	1
Total	$0 \leq$ score ≤ 9

Table 3.1: CHA2DS2VASc score criteria (from [167]).

Computing the value of each criterion and adding these values into the global CHA2DS2VASc score can both be achieved with different combinations of Java, OWL-DL 2.0 and SWRL. First, accessing the value of each criterion associated to a patient requires to follow several properties (typically *dolce:has-quality* from the patient to the CHA2DS2VASc score, then *dolce:has-quale* from the score to each criterion, then *has-integer-value* from each criterion to its value). Following the properties can be done either explicitly, or by using OWL-DL property chains. Second, some criteria values such as gender or the age thresholds can either be computed by a rather simple Java function, or using the ontology, which in turn can be achieved using OWL-DL features (e.g., a necessary and sufficient condition on a datatype property for the age) or SWRL (using built-ins). Third, adding the criteria values to compute the global CHA2DS2VASc score can also be performed by a Java function or by the `swrlb:add()` SWRL built-in. Some choice for one of the three previous steps may exclude some other choices in another step. For example, using Java to compute the age and gender criteria values only makes sense if the addition of the eight criteria values is itself done in Java. We systematically combined the Java, OWL and SWRL features for the three previous steps and derived ten possible strategies (cf. section 3.4.4 on the following page).

For each CHA2DS2VASc criterion, we manually determined whether reference to domain knowledge was necessary to reconcile the granularity gap with patient data (Table 3.2 on the next page).

We generated a validation set of 192 dummy patients representing all the combinations of values for the CHA2DS2VASc criteria. We validated each strategy by having it compute the CHA2DS2VASc score of each patient and comparing it to the solution. The solution is straightforward to compute separately by a program because when creating each patient from the validation set, the value of each criterion is known. For each strategy, we also measured the number of OWL classes, properties and SWRL rules used, as well as the CPU usage for each dummy patient. For each strategy, we measured the computing time on the evaluation set 50 times. Measurements were performed on a Dell Precision T3400 workstation with an Intel core 2 quad Q6600 64 bits 2,4 GHZ processor, 4 Gb RAM and Hitachi Ultrastar disk

criterion	ontology required	data example
Congestive heart failure	yes	"Diastolic heart failure", subclass of "Congestive heart failure" (patient 52)
Hypertension	yes	
Age	no	
Diabetes mellitus	yes	"Type-2 diabetes", subclass of "Diabetes mellitus" (patient 72)
Stroke	yes	"Ischemic stroke", subclass of "Stroke" (patient 57)
Vascular disease	yes	"Lower extremity occlusive peripheral heart disease", subclass of "Peripheral artery disease", which is in turn subclass of "Vascular disease" (patient 15)
$65 \leq$ **A**ge < 75	no	
Sex **c**ategory	no	

Table 3.2: Ontology requirement for computing the value of CHA2DS2VASc criteria. Patient numbers in the "Example" column refer to the evaluation set patients.

15K300 (300 Gb 15000 RPM). The test program was developed in Java on a Linux Ubuntu 11.04 (kernel 2.6.35-24) machine with open jdk 1.6.0_20, OWL API 3.2.0[8] and Pellet reasoner[9] API 2.2.1 [171].

We generated an evaluation set of patients implanted with an ICD and having an atrial fibrillation alert from the Paradym cohort[10]. Out of 74 patients, we selected the 62 patients having at least one document. We automatically calculated their CHA2DS2VASc score and recorded the performances of the ten strategies. We repeated the measurement 50 times. A physician (AB) manually calculated the reference CHA2DS2VASc score for each patient, and the result was double-checked by a cardiologist electrophysiologist (AR) for complex cases.

3.4.4 Modeling strategies

The systematic combination of Java, OWL and SWRL features to (i) access the value of each criterion, (ii) determine the value of some of the criteria and (iii) add the criteria values resulted in ten possible modeling strategies. Figure 3.19 on the facing page presents the decision tree explaining how each strategy was derived.

Some of the strategies can be constructed by adding defined classes or SWRL rules to other strategies (cf. original article [92] for details). Figure 3.20 on page 78 illustrates the dependencies between the various strategies.

Table 3.3 on the facing page illustrates the 10 strategies complexity by presenting their number of classes, properties and SWRL rules.

[8] `http://owlapi.sourceforge.net/`

[9] `http://clarkparsia.com/pellet`

[10] `http://clinicaltrials.gov/ct2/show/NCT01169246`

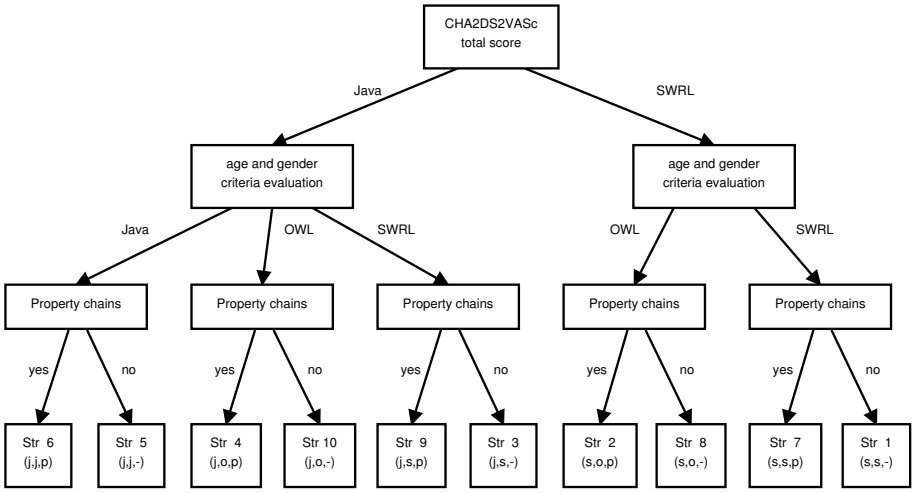

Figure 3.19: Decision tree combining the 10 possible strategies for modeling the CHA2DS2VASc criteria.

Strategy	Classes	Obj. prop.	datatype prop.	SWRL rules
1	0	0	3	12
2	2	1	0	9
3	0	0	0	8
4	2	1	0	8
5	0	0	0	5
6	0	1	0	5
7	0	1	0	9
8	2	0	3	12
9	0	1	0	8
10	2	0	0	8

Table 3.3: Complexity of the ten strategies in terms of number of additional classes, properties and SWRL rules over the base `patient.owl` model.

3.4.5 Comparison of the strategies' performances

On the validation set, all ten strategies computed the correct CHA2DS2VASc score for each 192 possible combination of criteria values. For each strategy, we computed the total time to process the 192 patients from the validation set. We then divided this total by 192 to determine the average computation time by patient so that it can be compared with the measures on the evaluation set, which is smaller. We repeated the operation 50 times. Figure 3.21 on page 79 shows the average computation time of a patient's CHA2DS2VASc score for each strategy.

On the evaluation set, all ten strategies computed the correct CHA2DS2VASc score for each of the 62 patients. Figure 3.22 on page 79 shows the average computation time of a patient's CHA2DS2VASc score for each strategy. Figure 3.23 on page 80 compares the average performance of each modeling strategy over the validation and evaluation sets.

All strategies computed the correct CHA2DS2VASc score for all the patients from the validation and evaluation sets. This demonstrates that both OWL and SWRL-based ontology modeling techniques can reliably perform the reasoning necessary to propose a severity level associated with ICD alerts.

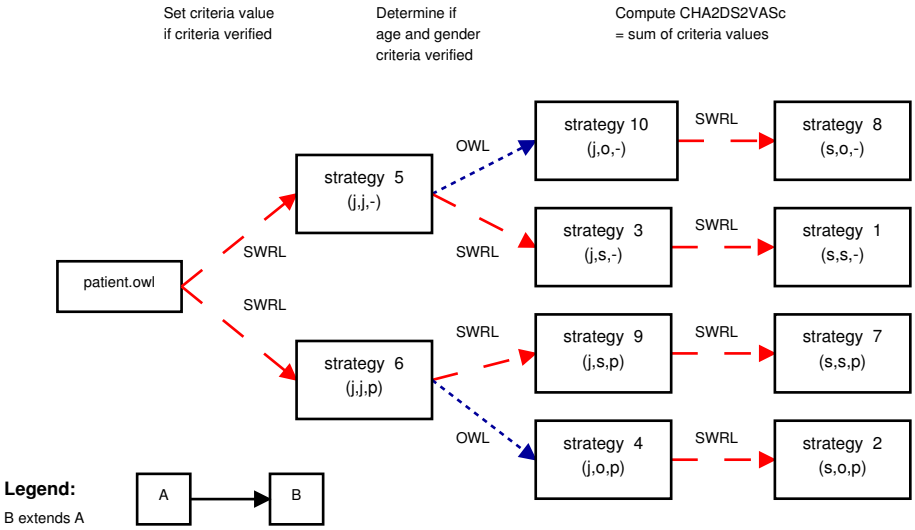

Figure 3.20: Dependencies between the common model (patient.owl) and the various strategies.

Table 3.3 on the preceding page and Figure 3.23 on page 80 show that the number of classes, properties and rules is not a good prediction of a strategy's performance, so this systematic study was relevant.

The comparison of the ten strategies showed that the best performances were not obtained by using exclusively OWL nor SWRL but by combining their respective advantages, using OWL to reduce the number of SWRL rules and making them simpler. Figure 3.23 on page 80 shows that the ranking of the strategies according to their performance is identical on the validation and evaluation set. For each strategy, the performance on the validation set was always better than on the evaluation set. This can be explained by the fact that there was no granularity-related reasoning involved in the validation set, whereas some patients from the evaluation set were associated with data more precise than the CHA2DS2VASc criteria. For example, patient 72 from the evaluation set had type 2 diabetes; a similar patient in the validation set would have been described as having diabetes. Another factor explaining the difference could be that the distributions of patients for each CHA2DS2VASc score were different for the validation and evaluation sets (cf. original article).

The modeling approach presented in this article potentially under-estimates a patient's CHA2DS2VASc score. If no information concerning a criterion is available, the criterion was assigned the value 0. We could have used a dual approach that over-estimates the score by assigning an initial value of 1 or 2 to each criterion and then setting it to 0 when there is some evidence that the criterion is not met. However, this would have had several drawbacks. First, clinical records typically mention what the patient has, and seldom mention what he has not, so except for the age and gender criteria, this lack of explicit information would result in assuming that almost all the criteria are met for all the patients. Second, the risk of thrombo-embolism is low for a CHA2DS2VASc score of 0, moderate for a score of 1 and high for a score between 2 and 9. The previous point would then result in important false positives, with almost all the patients being associated with a high risk. Eventually, combining the two approaches would provide an interval of validity for the CHA2DS2VASc score.

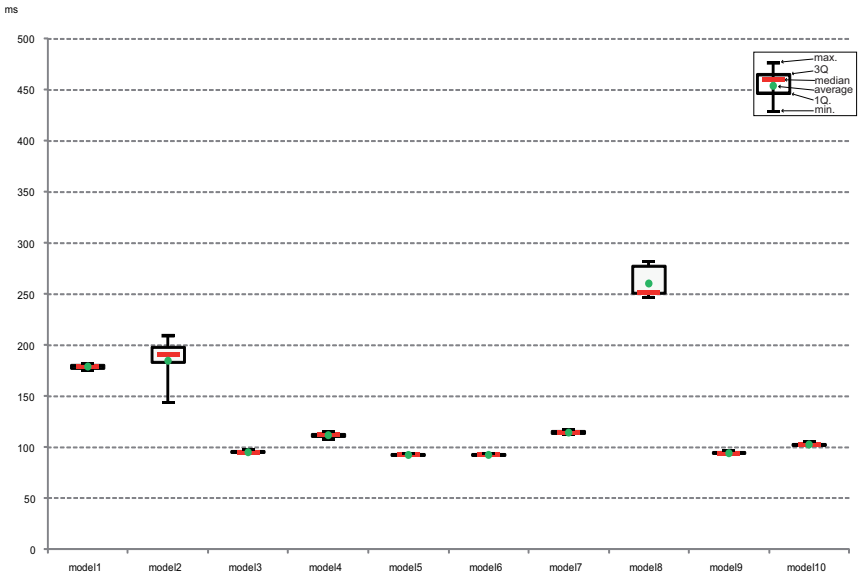

Figure 3.21: Boxplots representing the average computation time of the CHA2DS2VASc score of patients from the validation set. The boxplots were generated by repeating the measure 50 times.

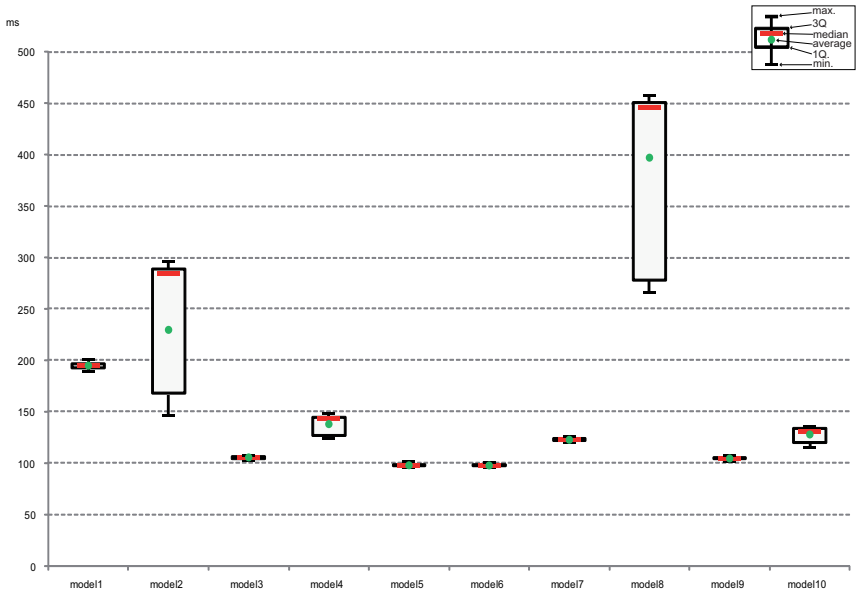

Figure 3.22: Boxplots representing the average computation time of the CHA2DS2VASc score of patients from the evaluation set. The boxplots were generated by repeating the measure 50 times.

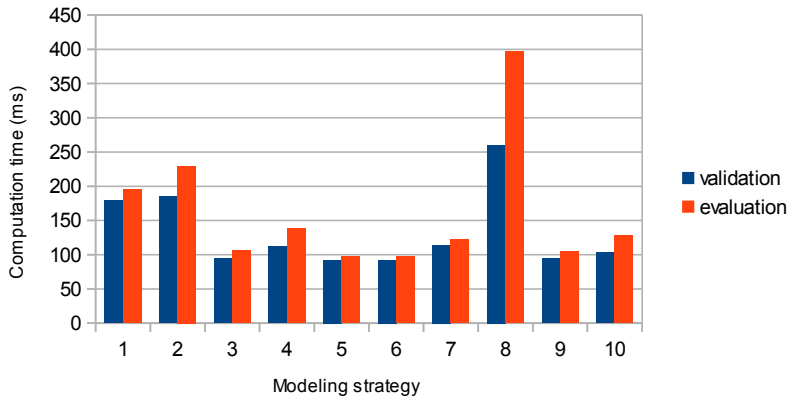

Figure 3.23: Average computation time of the CHA2DS2VASc score of the patients from the validation and evaluation sets.

3.5 Synthesis

What the three studies have in common In all three studies, we found that the lack of semantically-rich ontologies was a major limitation for building applications that involve reasoning more complex than retrieving the ancestors or the descendants of a class. Even if most biomedical ontologies from repositories such as BioPortal are available in OWL format, they are basically mere RDFS taxonomies. Few of them contain disjointness, existential or universal constraints, or even (correctly) defined classes as we will see in section 4.2.3.

This is all the more regrettable that all three studies also repeatedly demonstrated that once semantically-rich ontologies are available, developing the application-specific part of the reasoning only required the addition or the modification of a limited number of classes and could be done in a matter of hours.

Representation of anatomy The effort for converting the FMA in OWL was part of a subsequent larger research effort involving several teams [156, 158, 172]. Overall, this experiment triggers several observations. First it confirmed that the OWL language was well designed, as all its constructs found an application for modeling anatomy. Second and somehow paradoxically, it showed that the additional effort required for using more expressive constraints contributed to make the job easier as it supported using design patterns and integrity constraints. Third, new initiatives such as SPIN and ShEx offer new research perspectives for representing these constraints and using them during ontology development and maintenance.

Diagnosis of injuries Using the anatomy ontology to perform some automatic inference of indirect injuries showed that provided the semantically-rich ontologies are available, developing some application-specific reasoning can be quite simple.

It was also a first attempt at comparing class-based and instance-based reasoning for a more principled modeling approach. It showed that both worked equally well. Instance-based reasoning was more appealing from a modeling point of view but was also more difficult to implement whenever closure were required.

This work is also interesting because it marked the transition to reasoning with incomplete information, which will be covered in chapter 4.

Modeling strategies of pacemaker alerts severity This study compared OWL and SWRL-based reasoning for classifying pacemaker alerts. It showed that the optimal modeling strategy combined features of OWL and of SWRL. Nowadays, this study should also cover SPARQL and SPIN[11].

As we mentioned, our approach potentially under-estimates a patient's CHA2DS2VASc score. Conversely, assuming the worse case scenario and decrementing the CHA2DS2VASc score when we find evidence that a criterion is not met would over-estimate the score. Combining both approaches would provide a confidence interval. This typically raises the problem of reasoning with incomplete information, which will be further discussed in chapter 4.

What we learned

- Formalizing knowledge is difficult.
- It is useful for maintaining large ontologies; we have seen it for the FMA, and the Gene Ontology Next Generation is another example.
- It is useful for performing rich queries [173].

This lack of available semantically-rich ontologies can be partially explained by the fact that to a certain extent, ontology editing tools such as Protégé or TopBraid and the associated tutorials have succeeded too well. Domain experts are able to create and maintain their ontologies (which is good) with a minimal understanding of knowledge representation principles and of Semantic Web technologies. The latter point would not be so bad if people from the knowledge representation community (myself included) had not been less and less involved in ontology development over the last decade. I observed this trend in all the research teams I have been involved in since my PhD (included), as more recently when I contributed to the development of the ATOL livestock ontology [174, 175, 176]. Contributing to ontology development is highly time consuming, and extremely difficult to convert into high impact publications. Moreover, the large part of craftsmanship and informatics skills involved in the design patterns, the consistency constraints and grasping DL are understandably perceived as too difficult by the domain experts. This is strikingly the case for the Gene Ontology consortium that for years keep on using their idiosyncratic OBO formalism which is suboptimal in terms of interoperability, expressivity, maintenance and reasoning [177], and that chose to ignore efforts such as the OWL-based Gene Ontology Next Generation [52]. I have the feeling that when ontologies gained a mainstream status in life science in the second half of the 2000's (cf. Figure 1.1 on page 13), the knowledge representation community failed to create a pool of "knowledge representation engineers" who would have been recognized as key partners and could have taken over when the "knowledge representation researchers" gradually shifted their research efforts. Now these engineers are sorely missed and even if their input is valuable, few perceive it as such. The data deluge may be our next opportunity to patch things up if we (as researchers) both succeed in developing in time sound data management plans for E-Science (cf. my research perspectives in section 6.1.1.2), and if we succeed to avoid the previous mistake by having these plans adopted by the life science community. Ontologies will probably be key components for handling the large quantities of data and metadata. I am eager to see whether these will be mere taxonomies or semantically-rich ontologies.

[11] `http://spinrdf.org/`

Chapter 4

Reasoning with incomplete information

Outline

This chapter elaborates on situations we have encountered earlier where an incomplete description lead to imprecise or biased results. In the Virtual Project, we had intentionally failed to provide a closure axiom for some anatomical entities in oder to demonstrate that (cf. section 3.3.3.3 on page 67). These entities were correctly inferred to be ischemic when necessary, but the system could not decide whether they were partially or totally ischemic. The Virtual Soldier project demonstrated that the reasoning system can gracefully handle missing information by using more general superclasses as a kind of "degraded mode".

Subsequent works went one step further and sought to restrict the set of solutions by focusing on the things that can be inferred to be false when no conclusion can be reached concerning the things that can be inferred to be true. When the description of the world is exhaustive, the reasoner will always be able to recognize the situations where a condition is satisfied. When the description of the world is incomplete however, we may not be able to make the distinction between the situations where we do not know whether the condition is true, and the situations where we know that the condition is not true. Therefore, the general idea was to generate automatically negated versions of conditions of interest, with necessary and sufficient definitions referring to the original condition, and then to check whether data are classified as instance or as subclasses of the conditions or of their negated version.

Section 4.2 shows how this principle was first applied to the problem of grading tumors. A tumor grade is either 1, 2, 3 or 4, so each grade was a distinct condition. In case of incomplete information, we may not be able to reach a conclusion as to which grade qualifies, so none of them is proposed. However, even if it is incomplete, the available information may be sufficient to rule out some of the grades, and we can therefore narrow the possible solutions.

With tumor grades, the only constraints were that the conditions are mutually-exclusive and that each tumor has a grade (even if we do not know which one). Section 4.3 focuses on determining the clinical trials a patient may be eligible to. In this case, conditions were the trials eligibility criteria and a patient's eligibility is defined by a conjunction of criteria or their negation, so the overall outcome is more complex to infer. We showed that not taking incomplete information into account leads to over-estimating patients rejection, and we proposed a design pattern for modeling clinical trial that addresses this issue.

4.1 Principle

As we have seen in section 3.1, OWL constraints must hold for any interpretation function. Therefore, during knowledge modeling, one not only has to specify the constraints that must be met (e.g. a hand has to have a thumb, an index, etc.), but also the relations that cannot exist (e.g. a heart can never be a part of the hand). As we have seen in section 3.2.3.4, this is typically done with closure constraints representing "the only possible values for this property must be instances of these classes".

Although the open world assumption imposes an additional burden, it has two major benefits over the closed world assumption [28]. First, it supports a finer description, with the possibility to distinguish mandatory values from optional values. For example, an individual hand may not have a thumb in case of amputation or of abnormality, or it may even have additional fingers in case of polydactily, but it has to have exactly one palm. Second, as we will see throughout this chapter, it supports correct reasoning even if the domain is not described exhaustively.

4.2 Methodology: grading tumors

This study focuses on determining the grade of brain tumors. This was motivated by the need to reuse the patients data from one study in order to compare them with data from another study, with the two studies relying on slightly different grading systems. We applied the classification techniques seen in the previous chapter to automate the process. However, as often, the patients data were sometimes incomplete, which introduced a bias if we use closed world reasoning, and lead OWL-based reasoning unable to propose any grade for these patients. We proposed a method based on the logical negation of each tumor grade in order to determine the grades that were incompatible with what we knew of the patients, thus narrowing the set of possible grades.
In retrospect, this work is interesting because in addition to another example of how semantically-rich ontologies can be reused by applications, it shows that semantically-rich ontologies can handle gracefully incomplete data, which are ubiquitous in life sciences.

This study was originally published in: Gwenaëlle Marquet, Olivier Dameron, Stephan Saikali, Jean Mosser, and Anita Burgun. Grading glioma tumors using OWL-DL and NCI thesaurus. In *Proceedings of the American Medical Informatics Association Conference AMIA'07*, pages 508–512, 2007 [79]. It elaborates on a previous work on grading lung tumors: Olivier Dameron, Élodie Roques, Daniel L. Rubin, Gwenaëlle Marquet, and Anita Burgun. Grading lung tumors using OWL-DL based reasoning. In *9th International Protégé Conference*, 2006 [78].

4.2.1 Context

Brain tumors represent 2.4 percent of all cancer deaths. Among tumor variables, tumor grade and histology appear to have the greatest effect on survival. Glioblastoma, with median survival shorter than twelve months, is a highly malignant (grade IV) glioma, which has the propensity to infiltrate throughout the brain in contrast to pilocytic astrocytoma of the posterior fossa, which does not spread and can be cured by surgery [178]. Traditionally, the grading (classification) of a tumor is determined by the evaluation of tumor characteristics by a pathologist.

The process of determining the grade of a tumor consists in checking if it meets a set of requirements. There are numerous systems for grading the glioma tumors. The reference

grading system is the World Health Organization (WHO) grading system [179]. This system assigns a grade from 1 to 4 to glioma, grade 1 being the least aggressive and grade 4 being the most aggressive. This classification is based on five histopathology criteria that are related to the degree of anaplasia: cellular density, nuclear atypia, mitosis, endothelial proliferation and necrosis. The WHO malignant grades are described as follows:

- WHO Grade IV: **cellular density** high, **nuclear atypia** marked, high **mitotic activity**, **necrosis** present, **endothelial proliferation** present.
- WHO Grade III: **cellular density** increased, distinct **nuclear atypia**, **mitotic activity** marked, **necrosis** absent, **endothelial proliferation** absent.
- WHO Grade II: **cellular density** moderately increased, occasional **nuclear atypia**, **mitotic activity** absent or 1 mitosis, **necrosis** absent, **endothelial proliferation** absent.

Grading tumors is typically a classification task. The grading system requires domain knowledge in order to fill the granularity gap between the tumor descriptions and the grade descriptions. However, applications such as decision support for pathologists or integration of data graded using different systems require some formal representations of the grade definitions and the background knowledge. Such representations are typically achieved using ontologies. For the past years, a lot of biomedical ontologies have been developed including NCI thesaurus [153] (NCIT) a major resource in the cancer research domain. The NCIT provides descriptions for the brain tumors. It also has classes for the grades. However, those classes have neither proper descriptions nor definitions. Therefore, they can not be used for the automatic grading of tumors, which requires an explicit and formal representation.

4.2.2 Objective

The goal of this study is to show how the version of the NCIT in OWL (Web Ontology Language) can be extended to automatically perform classification of glioma using histological descriptions. We have focused our study on the malignant grade. For that, we have developed an ontology of the glioma tumors based on the World Health Organization grading system [179]. In this study, we focus on the reasoning tasks. We provide an overview of the TNM grading system. We then analyze the NCIT and conclude that it has to be extended in order to perform automatic tumor grading. We present the method that we used and the results obtained during the classification of a set of tests generated and the classification of eleven reports provided by a pathology department.

4.2.3 Why the NCIT is not up to the task

The NCI Thesaurus is a public domain Description Logic-based terminology to meet the needs of the cancer research community[153]. Its goal is to provide unambiguous codes and definitions for concepts used in cancer research. The NCIT has been converted into OWL-Lite [81]. The current version (07_01d) is composed of 55,458 named classes and 113 OWL properties. Among these classes, 18 % are defined classes, i.e. they have at least one necessary and sufficient constraint, and 82 % are primitive classes, i.e. they can have constraints, but do not have any necessary and sufficient definitions.

The classes representing the grades according to the WHO system have no restriction and are not semantically defined (Figure 4.1 on the next page). Therefore, they are just placeholders as nothing can be inferred to be a subclass or an instance of these classes. Because of the open-world-assumption underlying the OWL semantics, if the grade of a tumor cannot be unequivocally inferred, the tumor will not be classified under any grade. For example, tumors that

could be grade I or grade II tumors are not classified anywhere. There is no explicit difference between the grades the tumor belongs to (here I and II) and those it cannot belong to (here III and IV).

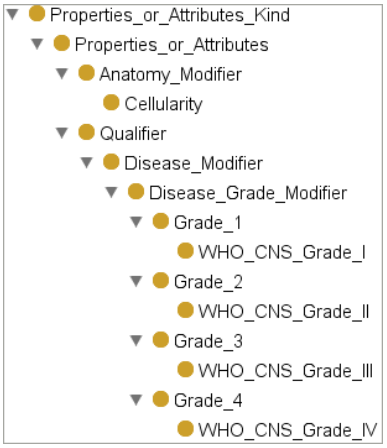

Figure 4.1: The WHO grades in the NCIT are primitive subclasses of `ncit:Disease_Grade_Modifier`. The intermediate classes `Grade_1` to `Grade_4` are placeholders allowing to take the multiple grading systems into account.

In the NCIT, the glioma tumors have been described as Central Nervous System Neoplasms. Each kind of tumors has been defined by necessary and sufficient conditions. For example, glioblastoma has been defined by the intersection of 17 restrictions. In Figure 4.2 on the facing page we present some conditions used to define the glioblastoma class.

Such definitions cannot be logically exploited to achieve any reasoning for several reasons. First, we see that being a grade 4 tumor is one of the conditions of the definition. Since the NCIT does not provide any definition for the grades, the grade cannot be inferred from the description of the tumor, which leaves to the user the task of stating the grade when describing the tumor... and if he knows the grade at this point, he probably does not need a reasoner to figure whether the tumor is a glioblastoma. Second, the constraint concerning the grade is a universal constraint ($\forall$), and again, leaves it to the user to make sure that the tumor is neither a grade 1, nor 2, nor 3. Moreover, such a restriction is difficult to represent with instances when describing a tumor. Third, the extensive use of "*`Disease_May_Have_...`* in existential constraints of the definition is deeply disturbing.

4.2.4 An ontology of glioblastoma based on the NCIT

The ontology we developed is based on the NCIT. A specific relevant part of the NCIT has been extracted using eleven terms corresponding to the names of the glioma tumors and nine terms that correspond to subclasses of atypia and mitotic activities. We first retrieved the NCIT classes corresponding to these terms and all their parents. For each of these classes, we followed all their relations and recursively retrieved the fillers and their parents.

Several operations have been necessary to address the issues mentioned in the previous sections and enhance the extracted portion of the NCIT. First we provided definitions for all WHO grades. Second, we added new classes (and new properties) for filling the granularity gap between the histologic features described in the WHO and the classes present in the NCIT. For handling the open-world-assumption, we also introduced the negations of each grade (namely nograde, cf. section 4.2.5 on page 89).

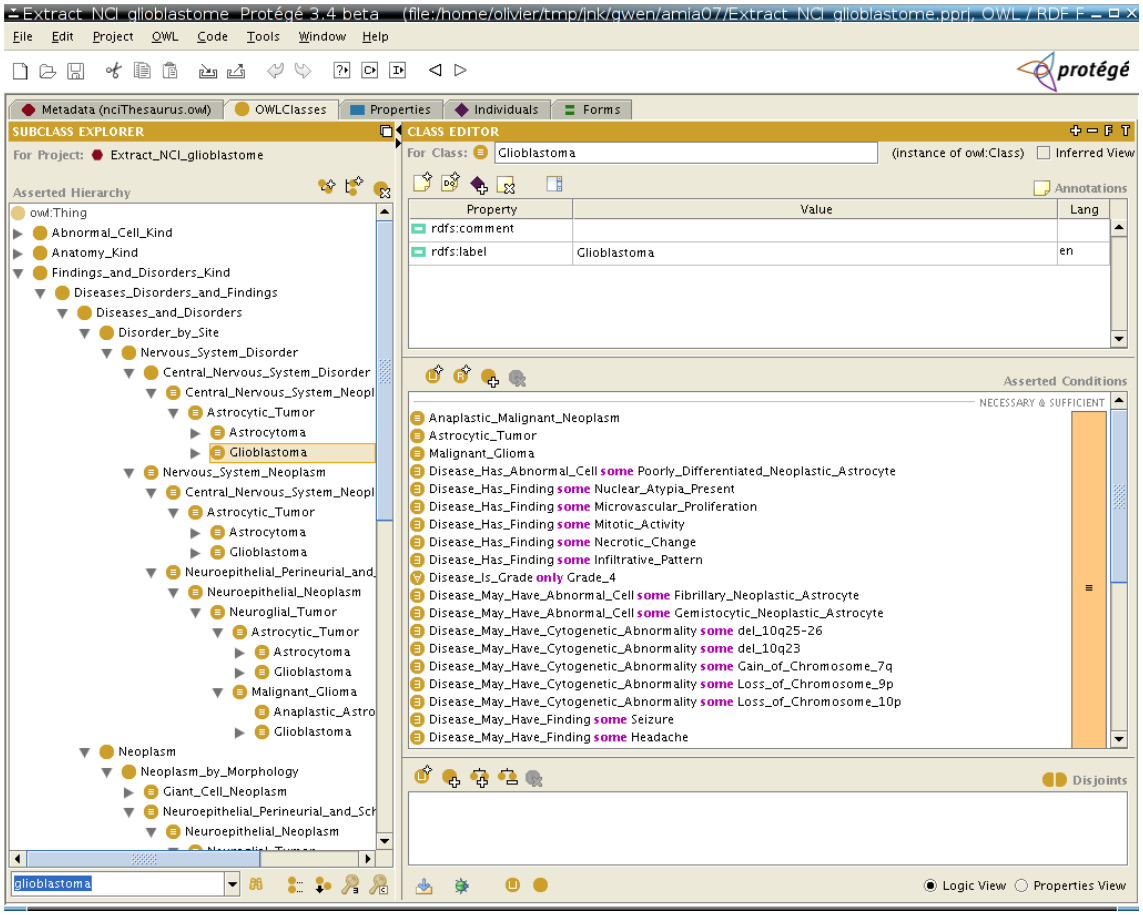

Figure 4.2: The `ncit:Glioblastoma` class has a necessary and sufficient definition. However, this definition cannot be logically exploited.

$$\texttt{WHO\_CNS\_GRADE\_II} \equiv \begin{cases} \texttt{nci:Disease\_Grade\_Modifier} \\ \sqcap \\ (\exists \textit{hasCellularDensity}\, \texttt{Moderate\_Increased\_Cellularity\_Present}) \\ \sqcap \\ (\exists \textit{hasAtypia}\, (\texttt{Occas.\_Nucl.\_Atypia\_Present} \sqcup \texttt{Dist.\_Nucl.\_Atypia\_Present})) \\ \sqcap \\ (\forall \textit{hasMitoticActivity}\, \texttt{Low\_Mitotic\_Activity}) \\ \sqcap \\ (\textit{hasNecrosisActivity} = 0) \\ \sqcap \\ (\textit{hasVascularProliferation} = 0) \end{cases}$$

$$\texttt{WHO\_CNS\_GRADE\_III} \equiv \begin{cases} \texttt{nci:Disease\_Grade\_Modifier} \\ \sqcap \\ (\exists \textit{hasCellularDensity}\, \texttt{Increased\_Cellularity\_Present}) \\ \sqcap \\ (\exists \textit{hasAtypia}\, (\texttt{Occas.\_Nucl.\_Atypia\_Present} \sqcup \texttt{Dist.\_Nucl.\_Atypia\_Present}) \\ \sqcap \\ (\exists \textit{hasMitoticActivity}\, \texttt{Marked\_Mitotic\_Activity}) \\ \sqcap \\ (\textit{hasNecrosisActivity} = 0) \\ \sqcap \\ (\textit{hasVascularProliferation} = 0) \end{cases}$$

$$\texttt{WHO\_CNS\_GRADE\_IV} \equiv \begin{cases} \texttt{nci:Disease\_Grade\_Modifier} \\ \sqcap \\ (\exists \textit{hasCellularDensity}\, \texttt{High\_Cellularity\_Present}) \\ \sqcap \\ (\exists \textit{hasAtypia}\, \texttt{Marked\_Nuclear\_Atypia\_Present}) \\ \sqcap \\ (\exists \textit{hasMitoticActivity}\, \texttt{High\_Mitotic\_Activity}) \\ \sqcap \\ (\exists \textit{hasNecrosisActivity}\, \texttt{Necrotic\_Change}) \\ \sqcap \\ (\exists \textit{hasVascularProliferation}\, \texttt{Vascular\_Proliferation}) \end{cases}$$

The generated ontology is composed of 243 classes, among which 33 are defined. Among the 243 classes, 234 classes correspond to NCIT classes, 5 classes have been added for the description of the histologic criteria and 4 classes have been added for the description of nogrades. We reused 24 class definitions from the NCIT and created the remaining 9.

Two sets of classification tests have been created. The validation set (15 tests) has been generated for representing plausible combinations of the histologic criteria. Each test corresponds to a prototypical tumor. The evaluation set corresponds to eleven pathologic reports provided by the pathology department of the Rennes hospital. Each report was represented as a subclass of `Disease_Grade_Modifier`. This step was performed manually. Each report is read and the corresponding Tumor class has been built manually. For each test, the description of its histologic criteria was done by existential restrictions for indicating the presence of a criterion, and

by cardinality restriction to zero for indicating the absence of a criterion. Figures 4.3 and 4.4 show the case of `Tumor10`, which is correctly inferred to be a grade IV tumor. All tumors from the validation set were correctly graded. Ten of the eleven tumors from the evaluation set were correctly graded. The remaining tumor's description only mentioned four of the five WHO criteria, so it was not classified as a subclass of any of the four grades.

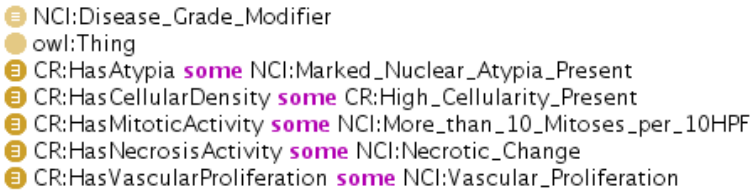

Figure 4.3: The `Tumor10` from the evaluation set is defined according to its characteristics.

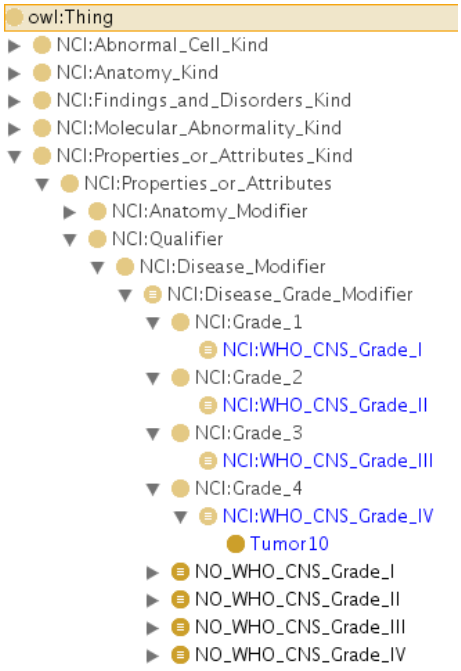

Figure 4.4: The `Tumor10` from the evaluation set is correctly inferred to be a grade IV tumor.

4.2.5 Narrowing the possible grades in case of incomplete information

We completed the ontology by making `WHO_CNS_GRADE_I`, `WHO_CNS_GRADE_II`, `WHO_CNS_GRADE_III` and `WHO_CNS_GRADE_IV` mutually-disjoint and by adding a coverage axiom to `Disease_Grade_Modifier`:

$$\texttt{Disease\_Grade\_Modifier} \equiv \begin{cases} \texttt{WHO\_CNS\_GRADE\_I} \\ \sqcup \\ \texttt{WHO\_CNS\_GRADE\_II} \\ \sqcup \\ \texttt{WHO\_CNS\_GRADE\_III} \\ \sqcup \\ \texttt{WHO\_CNS\_GRADE\_IV} \end{cases}$$

Eventually, we added the four `NoGrade` classes according to the template:

$$\texttt{NO\_WHO\_CNS\_GRADE\_I} \equiv \begin{cases} \texttt{Disease\_Grade\_Modifier} \\ \sqcap \\ \neg\texttt{WHO\_CNS\_GRADE\_I} \end{cases}$$

Figure 4.5 shows that `Tumeur4`, which grade could not be determined because of its incomplete description, was (correctly) classified as a subclass of both `NO_WHO_CNS_GRADE_III` and `NO_WHO_CNS_GRADE_IV`. This shows that even incomplete information can be valuable because it can be exploited to reduce the space of solutions. Here we could not infer the grade of the tumor, but the nograde classes allowed us to rule out grades 3 and 4 (the worse).

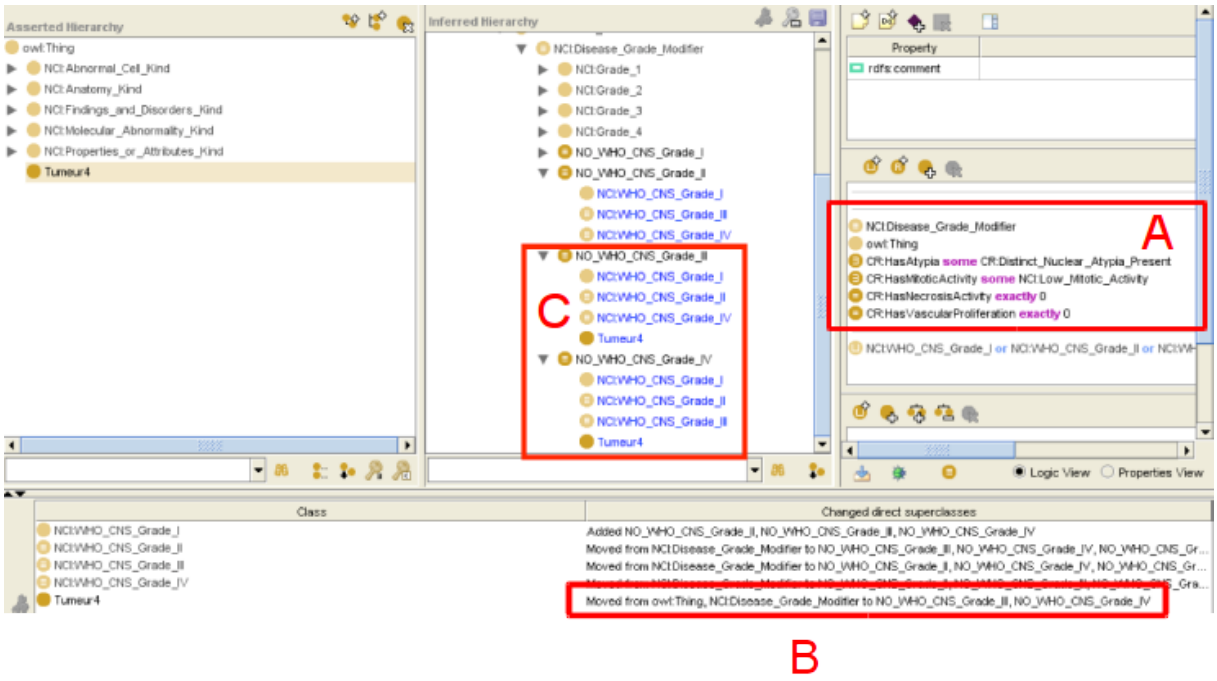

Figure 4.5: The `Tumor4` from the evaluation set only has a partial description that covers 4 out of 5 WHO grade criteria (A). This prevents the reasoner to infer its grade (B). However, the `NO_WHO_CNS_Grade` classes are useful for ruling out grades 3 and 4 (C).

4.3 Methodology: clinical trials recruitment

This study focuses on patient recruitment in clinical trials. This task requires the matching of a large volume of information about the patient with numerous eligibility criteria, in a logically-complex combination. Moreover, some of the patient's information necessary to determine the status of the eligibility criteria may not be available at the time of pre-screening. We showed that the classic approach based on negation as failure over-estimates rejection when confronted with partially-known information about the eligibility criteria because it ignores the distinction between a trial for which patient eligibility should be rejected and trials for which patient eligibility cannot be asserted. We have also shown that 58.64% of the values were unknown in the 286 prostate cancer cases examined during the weekly urology multidisciplinary meetings at Rennes' university hospital between October 2008 and March 2009. We proposed an OWL design pattern for modeling eligibility criteria based on the open world assumption to address the missing information problem.
In retrospect, this work is interesting because it shows that the approach we developed for the grade of brain tumors is actually more general and can be adapted to other situations.

This study was a contribution to the Astec[1] project (ANR-08-TECS-0002) and was related to the EHR4CR[2] project (IMI 115189). It was originally published in: Olivier Dameron, Paolo Besana, Oussama Zekri, Annabel Bourdé, Anita Burgun, and Marc Cuggia. OWL model of clinical trial eligibility criteria compatible with partially-known information. *Journal of Biomedical Semantics*, 4(1), 2013 [95].

4.3.1 Context

Patient recruitment is a major focus in all clinical trials. Adequate enrollment provides a base for projected participant retention, resulting in evaluative patient data. Identification of eligible patients for clinical trials (from the principal investigator's perspective) or identification of clinical trials in which the patient can be enrolled (from the patient's perspective) is an essential phase of clinical research and an active area of medical informatics research. The National Cancer Institute has identified several barriers that health care professionals claim in regard to clinical trial participation[3]. Among those barriers, lack of awareness of appropriate clinical trials is frequently mentioned.

Automated tools that help perform a systematic screening either of the potential clinical trials for a patient, or of the potential patients for a clinical trial could overcome this barrier [180]. The ASTEC (Automatic Selection of clinical Trials based on Eligibility Criteria) project aims at automating the search of prostate cancer clinical trials to which patients could be enrolled to [181]. It features syntactic and semantic interoperability between the oncologic electronic medical records and the recruitment decision system using a set of international standards (HL7 and NCIT), and the inference method is based on ERGO [182]. The EHR4CR project aims at facilitating clinical trial design and patient recruitment by developing tools and services that reuse data from heterogeneous electronic health records. The TRANSFoRm project has similar objectives for primary care [183, 184].

All these studies on data and criteria representation, integration and reasoning are motivated by the requirement to have the necessary information available at the time of processing

[1] `http://www.agence-nationale-recherche.fr/?Project=ANR-08-TECS-0002`
[2] `http://www.ehr4cr.eu/`
[3] `http://www.cancer.gov/clinicaltrials/learningabout/in-depth-program/page7`

the patient's data, and assume that somehow, that will be the case. Missing information that is required for deciding whether a criterion is met leads to recruitment being underestimated. Solutions for circumventing this difficulty consist either in making assumptions about the undecided criteria, or in having a pre-screening phase considering a subset of the criteria for which patient's data are assumed to be available. Bayesian belief networks have been used to address the former [185] but require a sensible choice of probability values and may lead to the wrong assumption in particular cases. The latter leaves most of the decision task to human expertise, which provides little added value (if an expert has to handle the difficult criteria, automatically processing the simple pre-screening ones is only a little weight off his shoulders) and is still susceptible to the problem of missing information for the pre-screening criteria.

4.3.2 Objective

We propose an OWL design pattern for modeling clinical trial eligibility criteria. This design pattern is based on the open world assumption for handling missing information. It infers whether a patient is eligible or not for a clinical trial, or if no definitive conclusion can be reached.

4.3.3 The problem of missing information

4.3.3.1 Modeling clinical trial eligibility

A clinical trial can be modeled as a pair $< (I_i)_{i=0}^{n}, (E_j)_{j=0}^{m} >$ where $(I_i)_{i=0}^{n}$ is the set of the inclusion criteria, and $(E_j)_{j=0}^{m}$ is the set of the exclusion criteria. All the eligibility criteria from $(I_i)_{i=0}^{n} \cup (E_j)_{j=0}^{m}$ are supposed to be independent from one another (at least in the weak sense: the value of criterion C_k cannot be inferred from the combined values of other criteria). Each criterion can be modeled as an unary predicate $C(p)$, where the variable p represents all the information available for the patient. $C(p)$ is true if and only if the criterion is met.

A patient is deemed eligible for a clinical trial if *all* the inclusion criteria and *none* of the exclusion criteria are met.

$$\boxed{\text{patient eligible} \Leftrightarrow \bigwedge_{i=0}^{n} I_i(p) \wedge \neg(\bigvee_{j=0}^{m} E_j(p))} \tag{4.1}$$

Before making the final decision on the list of clinical trials for which a patient is eligible for, there are intermediate pre-screening phases where only the main eligibility criteria of each clinical trial are considered. Such pre-screening sessions rely on subsets of $(I_i)_{i=0}^{n}$ and $(E_j)_{j=0}^{m}$, but the decision process remains the same.

For the sake of clarity, in addition to the general case, we will consider a simple clinical trial with two inclusion criteria I_0 and I_1, and two exclusion criteria E_0 and E_1.

$$\text{patient eligible} \Leftrightarrow I_0(p) \wedge I_1(p) \wedge \neg(E_0(p) \vee E_1(p)) \tag{4.2}$$

For example, these criteria could be:

- I_0: evidence of a prostate adenocarcinoma;
- I_1: absence of metastasis;
- E_0: patient older than 70 years old;
- E_1: evidence of diabetes.

According to equation 4.2, a patient would be eligible for the clinical trial if and only if he has a prostate adenocarcinoma and has no metastasis and is neither older than 70 years old nor suffers from diabetes.

Because of De Morgan's laws, equation 4.1 is equivalent to:

$$\boxed{\text{patient eligible} \Leftrightarrow (\bigwedge_{i=0}^{n} I_i(p)) \wedge (\bigwedge_{j=0}^{m} \neg E_j(p))} \tag{4.3}$$

Even though equation 4.1 and equation 4.3 are logically equivalent, the latter is often preferred because it is an uniform conjunction of criteria. Note that the negations in front of the exclusion criteria are purely formal, as both inclusion and exclusion criteria can represent an asserted presence (e.g. prostate adenocarcinoma for I_0 or of diabetes for E_1) or an asserted absence (e.g. metastasis for I_1).

For our example:

$$\text{patient eligible} \Leftrightarrow I_0(p) \wedge I_1(p) \wedge (\neg E_0(p)) \wedge (\neg E_1(p)) \tag{4.4}$$

According to equation 4.3, a patient would be eligible for the clinical trial if and only if he has a prostate adenocarcinoma and has no metastasis and is not older than 70 years old and does not suffer from diabetes.

4.3.3.2 Patients who we know are not eligible and those who we do not know whether they are eligible

When a part of the information necessary for determining if at least one criterion is met is unknown, the conjunction of equation 4.3 can never be true. This necessarily makes the patient not eligible for the clinical trial, whereas the correct interpretation of the situation is that the patient cannot be proven to be eligible. This is different from proving that the patient is not eligible, and indeed, in reality the patient can sometimes be included by assuming the missing values (cf. next section).

For our fictitious clinical trial, we consider a population of nine patients covering all the combinations of "*True*", "*False*" or "*Unknown*" for the inclusion criterion I_1 and the exclusion criterion E_1. Table 4.1 on the next page presents the value of equation 4.4 and correct inclusion decision for the nine combinations. Among the five patients (p_2, p_5, p_6, p_7 and p_8) for which at least a part of the information is unknown, three (p_2, p_7 and p_8) illustrate a conflict between the value of equation 4.4 and expected inclusion decision. A strict interpretation of equation 4.4 leads to the exclusion of the eight patients:

- for three of them (p_0, p_3 and p_4), all the information is available;
- for two of them (p_5 and p_6), some information is unknown, but the available information is sufficient to conclude that the patients are not eligible;
- for the three others (p_2, p_7 and p_8), however, the cause of rejection is either because one of the inclusion criteria cannot be proven (I_1 for p_7 and p_8) or because one of the exclusion criteria cannot be proven to be false (E_1 for p_2 and p_8).

In the case of unknown information, equation 4.3 alone is not enough to make the distinction between the patients we know are not eligible (the first two categories, so this also includes patients for whom a part of the information is unknown) and those we do not know if they are eligible (the third category). This is a problem because patients from the first two categories should be excluded from the clinical trial, whereas those from the third category should be considered for inclusion.

Patient	I_0	I_1	E_0	E_1	$I_0 \wedge I_1 \wedge \neg E_0 \wedge \neg E_1$	Decision
p_0	T	T	F	T	F	Exclude (E_1)
p_1	T	T	F	F	T	Include
p_2	T	T	F	?	**F** cannot assert $\neg E_1$	**Propose** (assume $\neg E_1$)
p_3	T	F	F	T	F	Exclude (both $\neg I_1$ and E_1)
p_4	T	F	F	F	F	Exclude ($\neg I_1$)
p_5	T	F	F	?	F	Exclude ($\neg I_1$)
p_6	T	?	F	T	F	Exclude (E_1)
p_7	T	?	F	F	**F** cannot assert I_1	**Propose** (assume I_1)
p_8	T	?	F	?	**F** cannot assert I_1 cannot assert $\neg E_1$	**Propose** (assume both I_1 and $\neg E_1$)

Table 4.1: Evaluation of equation 4.4 and correct inclusion decision for all the possible values of I_1 and E_1, with possibly unknown information.

One solution could be to assume the values of the unknown criteria in order to go back to a situation where inclusion or exclusion could be computed using equation 4.3. In this case:

- inclusion criteria for which the available information is not sufficient to compute the status are considered to be met;
- exclusion criteria for which the available information is not sufficient to compute the status are considered not to be met.

Therefore, in the case where the available information is not sufficient to compute the status of a criterion, a different status is assumed depending on whether the criterion determines inclusion or exclusion. Referring to our fictitious clinical trial, the lack of information about the absence of metastasis would lead to the assumption that I_1 is true, whereas the lack of information about diabetes would lead to the assumption that E_1 is false.

This situation raises several issues:

- a different status is assumed depending on whether the criterion determines inclusion or exclusion;
- the assumed status depends on the nature of the criterion (i.e. inclusion or exclusion) and not on its probability;
- one has to remember that the value for at least a criterion has been assumed in order to qualify the inferred eligibility (adamant for p_0 or p_1 vs "under the assumption that..." for p_2, p_7 and p_8);
- this qualification can be difficult to compute (the status of E_1 is unknown for both p_2 and p_5, but p_5 can be confidently excluded whereas p_2 can be included assuming E_1).

4.3.4 Eligibility criteria design pattern

- for each criterion, create a class `C_i` (at this point, we do not care if it is an inclusion or an exclusion criterion, or both) and possibly add a necessary and sufficient definition representing the criterion itself (or use SWRL);
- for each criterion, create a class `Not_C_i` defined as `Not_C_i` $\equiv$ `Criterion` $\sqcap\neg$ `C_i`.This process can be automated;
- for each clinical trial, create a class `Ct_k` (placeholder);
- for each clinical trial, create a class `Ct_k_include` as a subclass of `Ct_k` with a necessary and sufficient definition representing the conjunction of the inclusion criteria and of the exclusion criteria (cf. equation 4.3) (`Ct_k_include` $\equiv \bigsqcap_{i=0}^{n}$ `I_i` $\sqcap \bigsqcap_{j=0}^{m}$ `Not_E_j`);
- for each clinical trial, create a class `Ct_k_exclude` (placeholder) as a subclass of `Ct_k`;
- for each clinical trial, create a class `Ct_k_exclude_at_least_one_exclusion_criterion` as a subclass of `Ct_k_exclude` with a necessary and sufficient definition representing the disjunction of the exclusion criteria (`Ct_k_exclude_at_least_one_exclusion_criterion` $\equiv \bigsqcup_{j=0}^{m}$ `E_j`);

- for each clinical trial, create a class `Ct_k_exclude_at_least_one_failed_inclusion_criterion` as a subclass of `Ct_k_exclude` with a necessary and sufficient definition representing the disjunction of the negated inclusion criteria
(`Ct_k_exclude_at_least_one_failed_incl_criterion` $\equiv \bigsqcup_{i=0}^{n}$ `Not_I_i`);

- represent the patient's data with instances (Figure 4.6 and 4.7). For the sake of simplicity, we will make the patient an instance of as many `C_i` as we know he matches criteria, and as many `Not_C_j` classes as we know he does not match criteria, even if this is ontologically questionable (a patient is not an instance of a criterion). How the patient's data are reconciled with the criteria by making the patient an instance of the criteria is not specified here: it can be manually, or automatically with OWL necessary and sufficient definitions or SWRL rules for the `C_i` and `Not_C_j` classes.

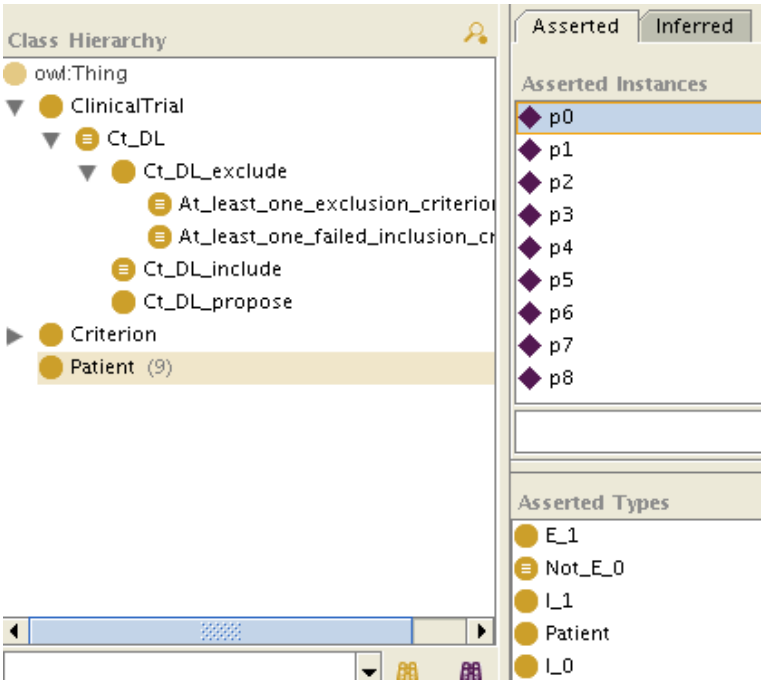

Figure 4.6: A patient for who all the information is available.

4.3.5 Reasoning

If all the required information is available, after classification, for each criterion the patient will be an instance of each `C_i` or `Not_C_i`, and therefore will also be instantiated as either `Ct_k_include` (like p_1 in Figure 4.8 on the facing page), `Ct_k_exclude_at_least_one_exclusion_criterion` or `Ct_k_exclude_at_least_one_failed_inclusion_criterion` (so at least we are doing as well as the other systems).

If not all the information is available, because of the open world assumption, there will be some criteria for which the patient will neither be classified as an instance of `C_i` nor of `Not_C_i` (e.g. in Figure 4.7 on the next page, p_2 is neither an instance of `E_1` nor of `Not_E_1`), so he will not be classified as an instance of `Ct_k_include` either. However, the patient may be classified

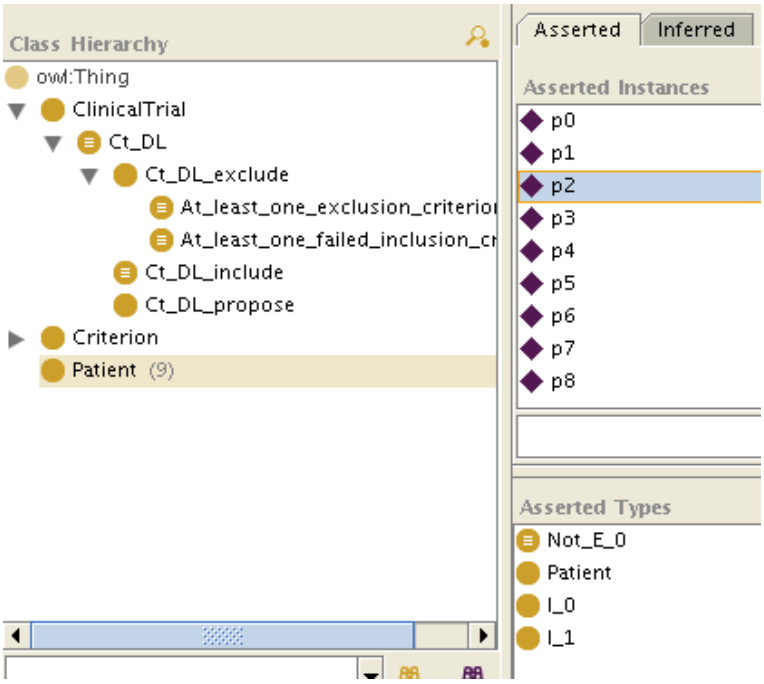

Figure 4.7: A patient for who some information is unknown (here about E_1).

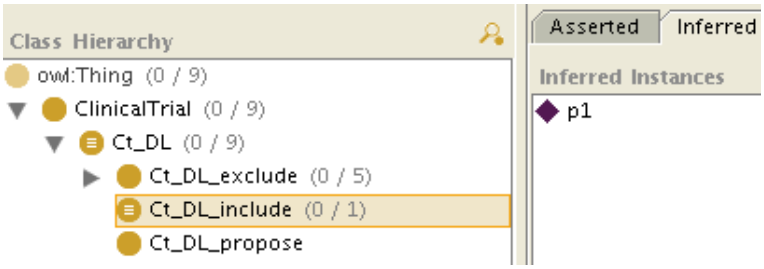

Figure 4.8: The class modeling clinical trial inclusion after classification (here patient p_1 can be included).

as an instance of
`Ct_k_exclude_at_least_one_exclusion_criterion` or of
`Ct_k_exclude_at_least_one_failed_inclusion_criterion`. As both are subclasses of `Ct_k_exclude`, we will conclude that the patient is not eligible for the clinical trial. We will even know if it is because he matched an exclusion criterion (like p_0, p_3 and p_6 in Figure 4.9), because he failed to match an inclusion criterion (like p_3, p_4 and p_5 in Figure 4.10), or both (like p_3).

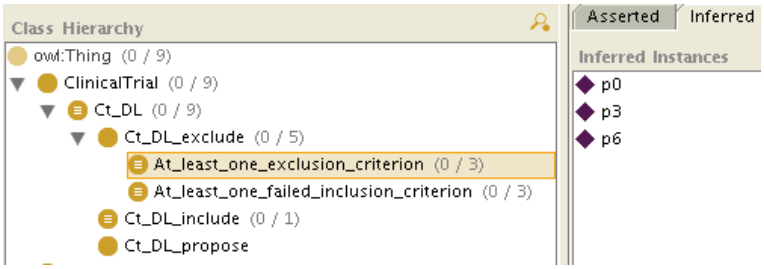

Figure 4.9: The class modeling clinical trial exclusion because at least one of the exclusion criteria has been met after classification (here patients p_0, p_3 and p_6 match the definition).

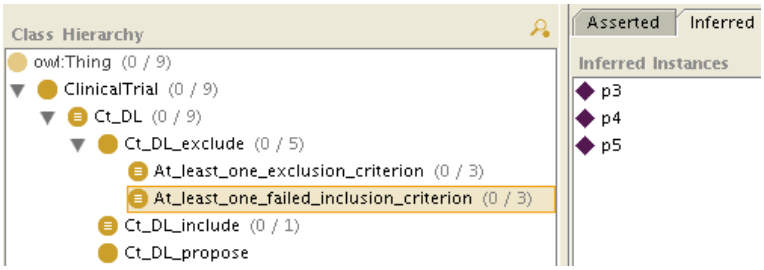

Figure 4.10: The class modeling clinical trial exclusion because at least one of the inclusion criteria failed to be met after classification (here patients p_3, p_4 and p_5 match the definition).

If the patient is neither classified as an instance of `Ct_k_include` nor of `Ct_k_exclude` (or its subclasses), then we will conclude that the patient can be considered for the clinical trial, assuming the missing information will not prevent it (like p_2, p_7 and p_8, who do not appear in Figs. 4.8, 4.9 and 4.10, consistently with Table 4.1 on page 94). By retrieving the criteria for which the patient is neither an instance of `C_i` nor of `Not_C_i`, we will know which information is missing.

4.4 Synthesis

What the two examples have in common The approach developed in this chapter and exemplified by the two reasoning applications could certainly be applied to many other contexts. Missing or incomplete information is pervasive in life sciences, and this is an inner characteristics. The study on clinical trials demonstrated the extent of the phenomenon, with about 68 % of patients data not specified, and none of the 286 patients having all the required fields for any of the four clinical trials we considered. I do not expect this trend to decrease, as we will keep on needing to combine patches of informations from different natures and different origins. However, part of what is causing the problem is also the solution: this integration endeavor is supported by new technologies that inherently take missing information into account. As

we are making the transition from the information silo paradigm to the linked data paradigm, we are switching from query languages such as SQL that rely on the closed-world assumption over well groomed and exhaustive data to reasoners capable of making the distinction between assertions that we know are false and assertions that are undecided (e.g. with "MINUS" and "NOT EXISTS" in SPARQL) or supporting the open-world assumption (OWL). I have the impression that these features are not exploited to their full potential yet, but insisting on using the former query languages on data that do not fulfill the exhaustivity requirement anymore does not seem sensible.

A conclusion of the chapter on reasoning based on classification was that modeling the domain-specific part of the reasoning task required a very little amount of work, provided that semantically-rich ontologies are available. Again, the availability of these ontologies turned out to be a limiting factor and we had to fix manually the imperfections of the NCIT.

Eventually, in both examples of grading tumors and of determining patients eligibility, the extra amount of work dedicated to handle missing information was again minimal. Moreover, both examples suggest that this additional work can be automated at least partially: the generation of the `NO_WHO_GRADE` classes and their definitions followed a simple pattern that was independent from the definitions of the grades. Similarly, the generation of the `Ct_exclude` and of the negated criteria classes as well as their definition were only formal manipulations relying on the definition of the corresponding `Ct_include`.

Grading tumors This application elaborates on a situation encountered in the Virtual Soldier project. It took me several years before I realized that what I considered then to be a concrete example of why closure axioms are useful in OWL (explaining the open world assumption in OWL was a major point during the Protégé Short Course and on the mailing list) was actually a part of a more general topic that proved to be valuable for modeling and reasoning over biomedical data.

Assessing patients eligibility to clinical trials This application is in turn a transposition of the tumor grading method. It could have implications in the more general efforts to formalize and standardize the representation of clinical trials eligibility criteria.

However, even if the solution I developed was adequate, I have since then replicated the reasoning mechanism using SPARQL instead of OWL. However, this mechanism only focuses on combining the status of the various eligibility criteria. Determining the status of each criterion typically remains a classification task for which OWL is best suited. Comparing the original solution with an approach relying on OWL for determining each criterion status and on SPARQL for combining the criteria remains to be done. It would be along the line of the optimization study from section 3.4.

What we learned

- Failing to explicitly address incomplete information may lead to biased results.
- The modeling overhead for taking incomplete information into account was marginal in both examples, and so was the additional computing cost.
- Just because some piece of information is incomplete does not mean that it is useless, as it can be exploited to reduce the space of solutions.
- The problem of incomplete information is pervasive in life science; however so far data sources and applications seldom take it into account, which make it a relevant field of

research. Primmer et al. made an in depth analysis of the Gene Ontology and its relevance for analyzing non-annotated genomes using what is known on model species [30]. It should be noted that Chen et al. developed a similarity measure among genes with shallow annotations [186]. Moreover, the Gene Ontology supports a `NOT` modifier for stating that a gene product was proved to be not associated with a GO term (e.g. for *Homo sapiens*, APOA1 (`uniprot:P02647`) is not associated to "transforming growth factor beta receptor signaling pathway" (`go:0007179`)). This modifier allows to make the distinction between the situations where we do not know whether a gene product is annotated to a GO term (absence of annotation) and the situations where we know that a gene product is not associated to the GO term (annotation with the `NOT` modifier). Even if such modifiers should be taken into account [140], I do not know of any widespread application that use them (which of course does not mean that there are no such applications).

- The question of confidence is also related to missing information. The Gene Ontology evidence codes[4] and the associated decision tree[5] were exploited by the IntelliGO semantic similarity measure [187]. GO evidence codes were later extended to the Evidence Ontology (ECO) [188] and inspired the Confidence Information Ontology [189].

[4] `http://geneontology.org/page/guide-go-evidence-codes`
[5] `http://geneontology.org/page/evidence-code-decision-tree`

Chapter 5

Reasoning with similarity and particularity

Outline

In addition to classification and deductive reasoning, life sciences data analysis also encompasses comparison. By making explicit the relations between classes, ontologies make it possible to go beyond simple annotation counting for determining what two elements have in common, or to what extent these two elements are different.

A collaboration with Christian Diot focused on the comparison of the lipid metabolism pathways for ducks and chicken. Ducks and geese produce *foie gras* when fattened whereas most other bird species produce abdominal fat instead, which lower the meat quality and its market value. Interestingly, *foie gras* is related to liver steatosis, a condition that can progress into fatty liver disease, cirrhosis or liver cancer in mammals and particularly humans. In this context, we supervised Charles Bettembourg's PhD thesis on a generic method based on semantics for the metabolic networks comparison across species. A major challenge was that most existing methods focus on what is similar, whereas we were specially interested in the differences. We proposed a method that first identifies the similar pathway steps and second identifies the similar steps associated to some specific processes in one of the species. This led us to define a semantic particularity measure as a complement to existing similarity measures (section 5.2), and to determine an objective discretization method for determining whether two elements were similar, and whether they are particular (section 5.3). The problem was further complicated by the fact that chicken or ducks are not as thoroughly annotated as human or mice. This bias rendered most of the classical similarity measures inadequate.

5.1 Principle

This section surveys the main categories of the numerous similarity measures and gives the definition of the measures used in sections 5.2 and 5.3.

The general principle consists in quantifying the similarity between two elements according to the annotations associated with each element. In certain domains, the process has also been extended for comparing two sets of elements. Similarity values usually range from 0 (low similarity) to 1 (perfect similarity).

Similarity is often seen as the dual notion of distance with the formula: $distance = 1 - similarity$, with distance values ranging from 0 (high similarity) to 1 (low similarity). However such distances are usually not proper distance metrics as they do not have the triangle inequality

property. Note also that the perspectives are different as similarity focuses on what is common between two elements, whereas distance focuses on what makes them different, so the connection between similarity and distance may not be straightforward and the previous formula should be seen as an approximation.

5.1.1 Comparing elements with independent annotations

5.1.1.1 Independent annotations with the same weight

Classic similarity measures are based on set operations over the annotations of two elements. If A and B are the sets of annotations of the first and the second element respectively, the Jaccard index is defined as :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

A similar notion is the Dice–Sørensen coefficient :

$$D(A,B) = \frac{2.|A \cap B|}{|A| + |B|}$$

There is a correspondence between the Jaccard index and the Dice–Sørensen coefficient :

$$\begin{cases} J = \dfrac{D}{2-D} \\ \\ D = \dfrac{2J}{1+J} \end{cases}$$

The Jaccard index and Dice–Sørensen coefficient both rely on two main assumptions: all the annotations have the same weight, and all the annotations have the same frequency. These two notions are different but not independent. Weight focus on the contribution of the annotation for determining the similarity between two elements. This is related to informativeness or granularity (a precise annotation conveys more information than a general or vague one). It is an intrinsic property of the annotation. Frequency is corpus-dependent, and is therefore an extrinsic property of annotations. Even if all the annotations had the same granularity, an annotation that annotates most of the elements of a corpus would be considered to be less informative than a rarer annotation. Of course, with annotations of different granularities, the more general annotations tend to be also the most frequent.

5.1.1.2 Independent annotations with different weights

The cosine similarity is a simple measure where the elements A and B to be compared are represented as vectors of n annotations. Each annotation has a fixed position in the vectors so that the i^{th} element of the vector of A refers to the same annotation as the i^{th} element of the vector of B.

$$similarity_{cosine}(A,B) = \frac{A.B}{||A|| \times ||B||} = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Although the i^{th} element of the annotation vector can be any real number (so cosine similarity is in the $[-1;1]$ range), it is usually a positive number (so cosine similarity is in the $[0;1]$ range). There are several classical strategies for determining the values of the annotation vector.

A binary vector representing the absence or the presence of annotations makes the cosine similarity applicable when all the annotations have the same weight (cf. section 5.1.1.1).

A more elaborate weighting scheme such as the "term frequency – inverse document frequency" (tf-idf) allows to take into account both the importance of the annotation (possibly different values for A and B), and the relative weights of each annotation (same value for A and B).

Term frequency indicates how important the annotation is to the element being compared. There are several weighting variants such as binary, raw frequency or log-normalization. For a text, this is typically the number of occurrences of a word divided by the number of words (for being able to compare texts of different lengths). For a gene, this is typically 1 or 0, depending on whether the gene is annotated or not. For a set of genes, this is typically the proportion of the genes in the set annotated by the term (for being able to compare sets of different sizes).

Inverse document frequency indicates how important the annotation is in general, according to a reference corpus. As mentioned previously, an annotation present in few documents is more informative than a common annotation. There are also several weighting variants such as the logarithm of the inverse frequency, i.e. the logarithm of the inverse of the proportion of documents in the corpus annotated by the term.

tf-idf is simply the product of the two previous aspects, which emphasizes an over-representation of rare annotations.

$$\begin{cases} tf(annotation, document) = \dfrac{Nb\ of\ occurrences\ of\ annotation\ in\ document}{Nb\ of\ annotations\ of\ document} \\ \\ idf(annotation, Corpus) = -log\dfrac{|\{d \in Corpus : annotation \in d\}|}{|Corpus|} \\ \\ tfidf(annotation, document, Corpus) = tf(annotation, document) \times idf(annotation, Corpus) \end{cases}$$

5.1.2 Taking the annotations underlying structure into account

All the previous similarity measures assume that the annotations are independent. However, the analysis can be further refined by using ontologies to also consider the relations between some of the annotations. Figure 5.1 on the next page presents the Gene Ontology hierarchy between three GO terms. This section shows how this hierarchy can be exploited by semantic similarity measures to infer that the first two are biologically close (their similarity is 0.57), whereas they are biologically different from the third (their similarity are respectively 0.08 and 0.11). Lee et al. performed a comparison of three families of similarity based respectively on IC, ontology structure and expert opinion on the SNOMED-CT ontology and found a poor agreement between IC-based metrics, whereas the metric based on ontology structure correlated best with expert opinion [190]. This suggests that taking the ontology structure into account improves the analysis, although whether this can be generalized to other ontologies and application contexts remains an open question.

Within a given gene set, the genes sharing identical or similar GO annotations can be grouped into clusters using two approaches [191]. The GSEA approach computes these clusters considering the GO terms over-representation. The semantic similarity approach takes into account GO properties to cluster genes considering the quantity and the importance of their shared annotations [192, 193, 194, 195]. Both approaches are not exclusive, as semantic measures can be involved in GSEA in order to improve the analysis [196]. If the GO terms were independent, the gene set characterization could be performed by a straightforward set-based

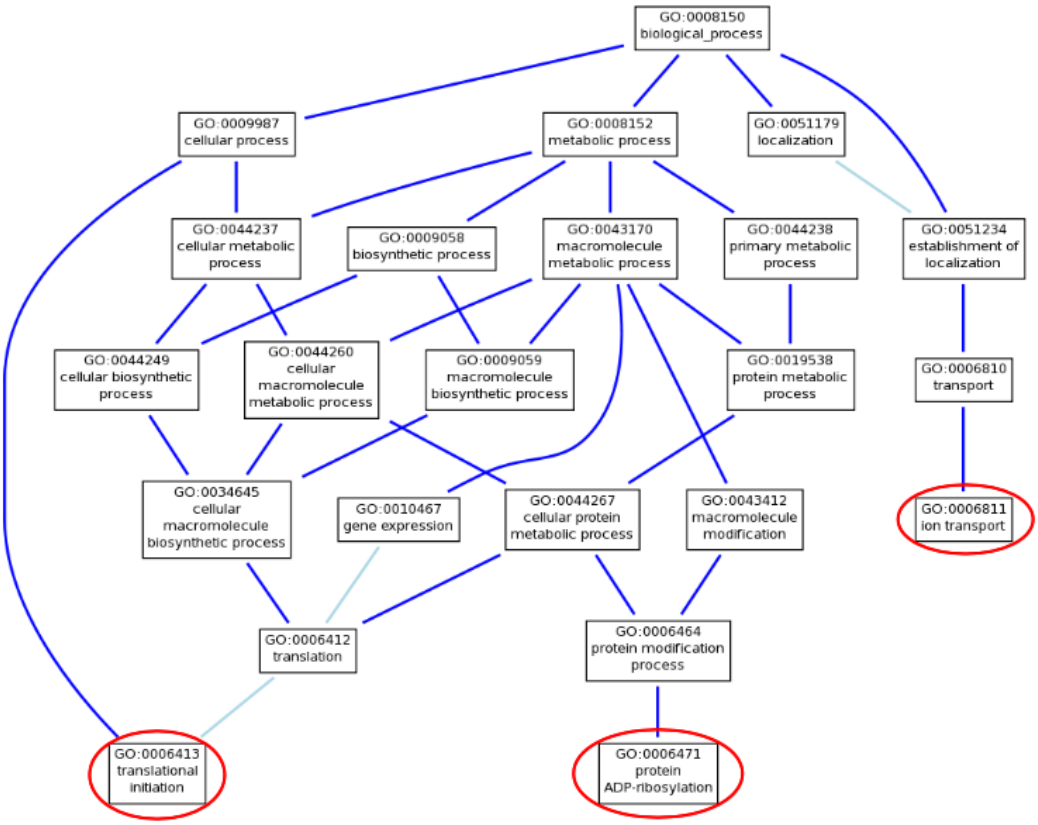

Figure 5.1: Gene Ontology hierarchy between "protein ADP-ribosylation", "translational initiation" and "ion transport". This hierarchy can be exploited by semantic similarity measures to infer that the first two are biologically close, whereas they are biologically further from the third. Dark blue edges represent "is_a" and light blue edges represent "part_of" relations (graph generated by Amigo).

approach such as the Jaccard index or Dice's coefficient. However, GO terms are hierarchically-linked. Consequently, the characterization needs to take into account the underlying ontological structure of the GO annotations [140].

Semantic similarity measures rely on ontologies to systematically quantify the weight of the shared elements. They exploit the formal representation of the meaning of the terms by considering the relations between the terms (e.g. for inferring new annotations that were implicit as each term inherits all the properties of its ancestors) and by attributing different weights to each term depending on how much information they convey. When working with annotation databases, it should be routine practice to use the ontology hierarchy to infer implicit annotation [140]. Gentleman developed a graph-based measure for the R package GOstats called simUI [197]. simUI defines the semantic similarity between two sets of terms corresponding to two sub-graphs of the ontology as the ratio of the number of terms in the intersection of those graphs to the number of GO terms in their union, which corresponds to a simple adaptation of the Jaccard index. However, with simUI, all the terms have the same weight, which introduces a bias emphasizing the intersection as the more general terms tend to annotate more genes. Other measures adopt different strategies to weight the terms. Pesquita et al. performed an extensive review of the main semantic similarity measures [198] and identified two main categories, i.e. node-based methods and edge-based methods, as well as a handful of hybrid methods. Blanchard et al. also performed an in-depth comparison of semantic similarities on subsumption hierarchies without multiple inference [199].

5.1.2.1 Node-based semantic similarity

Node-based semantic similarity measures rely on how informative the terms are. Typically, they consider that two terms sharing an informative lowest common ancestor are more similar than two terms with a less informative lowest common ancestor, as seen in Fig. 5.1 on the facing page.

Historically, Information Content (IC) value was used to quantify how informative a term is, with the least frequent terms having the highest IC value. Terms frequencies were determined using a reference corpus. The IC of a term t is its negative log probability $P(t)$. When the annotations are organized in an ontology such as GO, it is necessary to take into account the subsumption hierarchy when computing this frequency in order to also consider implicit annotations to the terms descendants [140].

$$IC(t) = -log(P(t))$$

This concept, borrowed from Shannon's Information Theory, was used to measure similarities using ontologies [200, 201, 202] such as WordNet [203]. To compare two terms, these methods rely on their most informative common ancestor (MICA). For Resnik, the similarity of two terms is simply the information content of their MICA. Lin also takes into account how far these two terms are from their MICA. Pesquita *et al.* proposed to combine the graph-based simUI metric with the IC of the terms involved in the computation [204]. In simGIC, each term is weighted by its IC.

$$similarity_{Resnik}(A, B) = max_{t \in (ancestors(A) \cap ancestors(B))}(IC(t))$$

$$similarity_{Lin}(A, B) = \frac{2 \times max_{t \in (ancestors(A) \cap ancestors(B))}(IC(t))}{IC(A) + IC(B)}$$

Ontologies are used twice when computing node-based semantic similarities: for determining the correct information content of annotations and for determining the most informative

common ancestor. These methods developed in linguistics have been applied to GO [205, 206] using the frequency with which a term annotates a gene as a marker of its rarity. Consequently, the IC of a GO term is inversely proportional to the frequency with which it annotates a gene using the Gene Ontology Annotations (GOA) database [138]. GOA specifies also how each annotation has be attributed through Evidence Codes (EC). In their method called "IntelliGO", Benabderrahmane et al. use a weighting corresponding to each GO term EC in addition to their IC [187].

Retrieving only the most informative common ancestor to compute a semantic similarity ignores the possibility that two GO terms can share several common ancestors. These situations result in a loss of information. A possible solution has been proposed that consists in using the average of the IC values of all disjoint common ancestors (DCA) instead of the maximum IC of this common set [207].

For the node-based methods relying on IC, the terms' frequencies used to compute the IC values depend on the corpus of reference. In the context of genes comparison, IC-based methods have three main limits related to their dependence on a GOA-based corpus. First, it can prove difficult or even impossible to obtain a relevant corpus. GOA provides single and multi-species annotation tables. Although using a species-specific table is well-suited to intra-species comparisons, it becomes problematic for cross-species comparisons. Second, using a multi-species table (like the UniprotKB table) in these cases is biased towards the most extensively annotated species such as human or mice. Third, the well-studied areas of biology have high annotation frequencies and are therefore less informative and see their importance downgraded, whereas the less-studied areas are artificially upgraded [208, 209, 210].

5.1.2.2 Edge-based semantic similarity

Edge-based semantic similarity measures use the directed graph topology to compute distances between the terms to compare. Among the simplest, Rada distance is based on the shortest path between the two terms [211], with extensions that rely on the average path among multiple paths [198]. Other approaches take into account the length of the path between the root of the ontology and the least common ancestor (LCA) of the terms, with the result that terms with a deep common ancestor are more similar than terms with a common ancestor close to the root [212, 213, 214, 215, 216]. The edge-based methods using depth as a proxy for precision are not dependent on a particular corpus. This can be a strength when it is difficult or impossible to determine a representative corpus, or a weakness when corpus-dependent frequencies are relevant. Moreover, another constraint to consider is that in most ontologies, granularity is not uniform so terms at the same depth can have different precisions; this is typically the case for GO [217].

5.1.2.3 Hybrid semantic similarity

Pesquita et al. also identified "hybrid" methods that combine different aspects of node-based and edge-based methods. In Wang's method [193], each term has a "semantic value" that represents how informative the term is, conforming to the node-based approach. However, the semantic value of a term is obtained by following the path from this term to the root and summing the semantic contributions of all the ancestors of this term. As the semantic value depends on the ontology topology, it also conforms to the edge-based approach. Note that this alternative approach is corpus-independent, so it is applicable when a relevant corpus cannot be computed (for comparing elements from several species) or does not exist (for poorly studied species). The relevance of the results obtained by this approach has previously been demonstrated [193, 198].

For computing a term's semantic value (SV), Wang first computes the semantic contributions of the ancestors of the term. In the following formulas, $S_A(t)$ is the semantic contribution of the term t to the term A and w_e is the semantic contribution factor for edge e linking a term t with its child term t'. According to Wang, we use a semantic contribution factor of 0.8 for the "is a" relations and 0.6 for the "part of" relations, and we added a 0.7 factor for the "[positively] [negatively] regulates" relations. An additional study not presented here showed that the value of the regulation factor had minimal impact (+/- 0.01) on the overall value.

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = max\{w_e * S_A(t') \mid t' \in children\ of\ (t)\}\ if\ t \neq A \end{cases}$$

Then, for each target term to compare, the semantic value is the sum of the semantic contributions of all its ancestors. Figure 5.2 shows an example of the semantic contributions of the ancestors of `GO:0043231` allowing to determine its semantic value: SV(`GO:0043231`) = 5.5952. The same operation for `GO:0005622` gives SV(`GO:0005622`) = 2.92 (Fig 5.3 on the following page). The more general a term (i.e. the less informative), the smaller its semantic value.

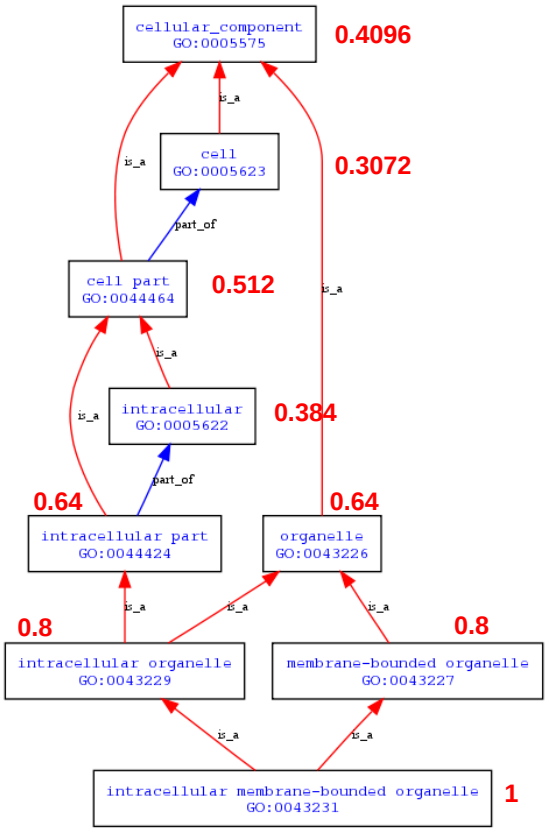

Figure 5.2: Semantic contributions of the ancestors of `GO:0043231`. The terms closer to `GO:0043231` contribute more. The farther the ancestor, the smaller its contribution to the term of interest. The semantic value of `GO:0043231` is the sum of its ancestors' semantic contribution,
here SV(`GO:0043231`) = 5.5952.

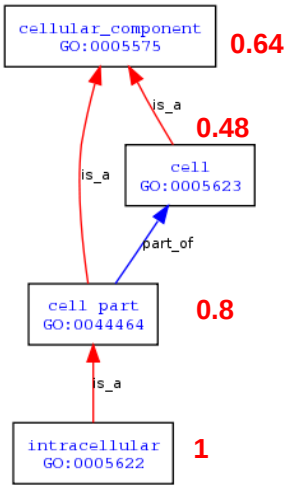

Figure 5.3: Semantic contributions of the ancestors of `GO:0005622`; here SV(`GO:0005622`) = 2.92.

$$SV(A) = \sum_{t \in T_A} S_A(t)$$

The terms semantic values and their ancestors' semantic contributions are used to compute the semantic similarity of two GO terms A and B:

$$similarity_{Wang}(A, B) = \frac{\sum_{t \in (T_A \cap T_B)} (S_A(t) + S_B(t))}{SV(A) + SV(B)}$$

The semantic similarity of a GO term A and a set of GO terms G is the highest similarity between A and each element of G:

$$similarity_{Wang}(A, G) = max_{t \in G}(similarity_{Wang}(A, t))$$

The semantic similarity of two sets of GO terms G_1 and G_2 is:

$$similarity_{Wang}(G_1, G_2) = \frac{(\sum_{t_1 \in G_1} similarity_{Wang}(t_1, G_2)) + (\sum_{t_2 \in G_2} similarity_{Wang}(t_2, G_1))}{|G_1| + |G_2|}$$

Pesquita et al. do not single out any particular semantic similarity measure as the best one, as the optimal measure will depend on the data to compare and the level of detail expected in the results. The main advantage of Wang's method compared to purely node-based methods is that the semantic value is not GOA-dependent, unlike information content. It is thus well-suited to cross-species comparisons. As cross-species comparison is one of the key stakes in biology, further development in the domain of semantic comparison should support such comparisons.

5.1.3 Synthesis

As we have seen, assessing the similarity of two elements can be greatly improved by using ontologies in order to take into account their annotations underlying structure. However, similarity alone is not enough for comparing biological pathways between non-model species. We

also need to identify the pathway steps that are similar between the two species but for which at least one of the two has some additional function. This sets up two challenges: being able to quantify similarity and particularity, and being able to determine both whether two elements are similar, and whether one of them has some particular function.

In the remainder of this chapter, section 5.2 presents a semantic particularity measure designed to be combined with any semantic similarity measure. The joint use of similarity and particularity allows to refine the comparison of sets based on the annotations of their elements. We show how the two sets similarity and their respective particularity determine comparison patterns (e.g. the two sets are similar and the second set presents a high particularity).

Section 5.3 presents a generic method for determining the optimal similarity and particularity thresholds minimizing the proportions of false positive and false negative as well as the abnormal comparison patterns.

5.2 Methodology: semantic particularity measure

This study focuses on the definition of a semantic particularity measure for comparing sets of elements annotated by an ontology. We propose to combine our particularity measure with a similarity measure to first identify the similar sets, and second identify sets with additional functions from among the similar ones. This particularity measure is initially applied to gene sets comparison according to the genes' GO annotations. We then show that the principle is generalizable to other ontologies.
In retrospect, this work is interesting because whereas semantic similarity has been an active research domain over the last decade with countless measures and not a single one outperforming the others, our approach allows to refine the analysis by also considering the specificities that are inherently ignored by similarity. Our semantic particularity measure is based on the general notion of informativeness, which can be derived from any semantic similarity measure, so combining similarity and the associated particularity can be performed with any similarity measure.

This study was originally published in: Charles Bettembourg, Christian Diot, and Olivier Dameron. Semantic particularity measure for functional characterization of gene sets using Gene Ontology. *PLoS ONE*, 9(1):e86525, 2014 [96].

5.2.1 Context

With the continued advance of high-throughput technologies, genetic and genomics data analyses are outputting large sets of genes. The amount of data involved requires automated comparison methods [4]. The characterization of these sets typically consists in a combination of the following three operations [218, 219]: first, synthesize the over- and under-represented functions of these genes [220, 221]; second, identify how these genes interact with each other [222]; third, identify and quantify the common shared features and the differentiating features [223, 224]. A widely used method for genes sets study called "Gene Set Enrichment Analysis" (GSEA) determines which gene features are over-represented in a gene set [225]. Numerous tools have been developed in this purpose: BiNGO [226], GOEAST [227], ClueGO [228], DAVID [229], GeneWeaver [230], GOTM [231]. See Hung et al. recent work for a review [232]. GSEA is useful for clustering a set of genes into subsets sharing over-represented features. Among these features, the biological processes (BP), molecular functions (MF) and cellular components (CC) annotating each gene are represented using the Gene Ontology (GO) [233]. GO is species-independent,

and thus supports cross-species comparison [30]. The GO graph itself is also widely used for genes semantic similarity analysis [234].

All the semantic similarity measures appear appropriate for identifying and quantifying common features. However, as these measures are focusing on common features, they may lead to an incomplete analysis when comparing genes having particular features along side similar ones [235]. For example, parts A and B of Figure 5.4 respectively present the molecular functions annotating the Exportin-5 orthologs of human (hsa) and rat (rno) and the Exportin-5 orthologs of human and drosophila (dme). Wang's method allows to compute cross-species semantic similarity. The results on MF annotations are: Sim(hsa, rno) = 0.797 and Sim(hsa, dme) = 0.726. This is consistent with the fact that globally, the Exportin-5 orthologs share the same functions between hsa, rno and dme. However, there are also five times as many human-specific MF terms compared to drosophila as compared to rats. It has been demonstrated that Exportin-5 orthologs are functionally divergent among species [236]. The tiny difference of semantic similarity (0.071) correctly reflects the fact that the orthologs share the same main function, but is not sufficient to identify that some species also have additional functions.

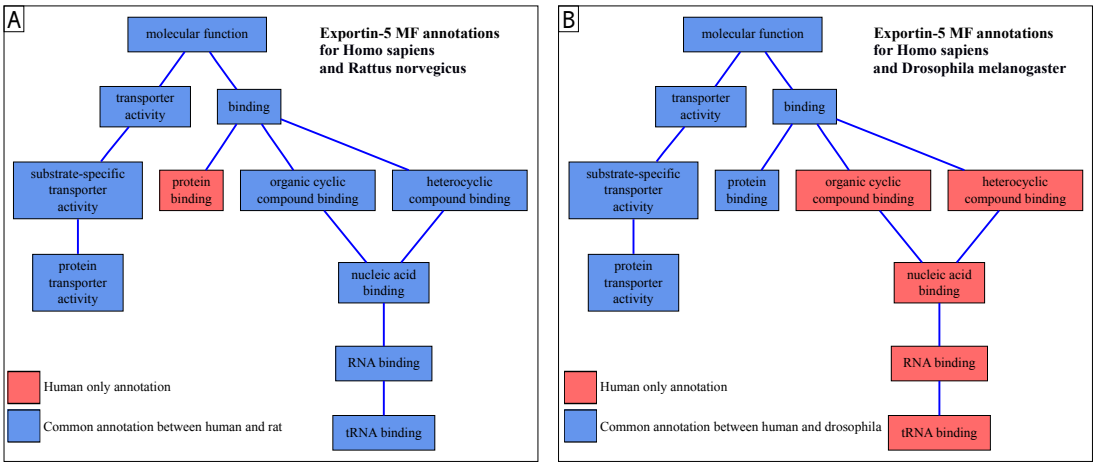

Figure 5.4: Representation of Exportin-5 orthologs annotations. Common terms between species are displayed in blue. The terms annotating only the human ortholog are displayed in red. Part A of this figure displays the MF annotations of the human and rat orthologs of Exportin-5. Part B displays the MF annotations of the human and drosophila orthologs of Exportin-5. In this example, there is no rat nor drosophila-specific term. The semantic similarity values obtained in these cases do not reflect the difference of human particularity between each part.

5.2.2 Objective

We assume that considering only similarity measures is not enough to compare sets of annotations. This analysis is valid for any set of annotations that refer to an ontology. We hypothesize that gene set analysis can be improved by considering gene particularities in addition to gene similarities. We propose a general definition and some associated formal properties. We propose also a new approach based on the notion of GO term informativeness to compute gene set particularities.

The original study was composed of three use cases. Section 5.2.6 on page 113 summarizes the second one.

- The first use case replicated Wang's study on *Saccharomyces cerevisiae* tryptophan degra-

dation when he defined his semantic similarity measure on GO. Our results showed that Wang's results are still valid. We also identified a benefit of using a particularity measure in addition to a similarity measure for identifying particular functions between similar genes.

- The second use case covered a larger dataset composed of 51 well annotated human genes related to aquaporin-mediated transport in order to determine whether similar genes with particular functions were a frequent situation. Our results showed that among similar genes, some also have some particular function and that this situation can be observed throughout the full range of similarity values.

- The third use case compared homolog genes across different species. Our results shows that ortholog genes were, as expected, mostly similar. Again, we also identified some of them having high particularity values that denote specific functions. Eventually, we identified some orthologs that have diverged and present a low similarity and high particularities.

5.2.3 Definition of semantic particularity

The semantic particularity of a set compared to another is the value that reflects the importance of the features that belong to the first set but not the second. To compare two genes, we rely on the similarity and the respective particularities of their sets of annotations. The particularity of a gene g1 annotated by the set Sg1 compared to a gene g2 annotated by the set Sg2 depends on the annotations of Sg1 that are not related to any annotation of Sg2.

5.2.4 Formal properties of semantic particularity

Like for semantic similarity, we compute a value bounded by 0 (least particular) and 1 (most particular). Four important properties arise from the semantic particularity definition:

- The semantic particularity is non-symmetric:

$$\text{Par}(\text{Sg1}, \text{Sg2}) = \text{x} \nRightarrow \text{Par}(\text{Sg2}, \text{Sg1}) = \text{x} \qquad \textbf{(Prop 1)}$$

- Compared to itself, a set of annotations has no semantic particularity:

$$\text{Par}(\text{Sg1}, \text{Sg1}) = 0 \qquad \textbf{(Prop 2)}$$

If $\text{Sg1} = \varnothing$, this comparison is meaningless.

- The semantic particularity of a set of annotations Sg1 ($\neq \varnothing$) is maximal when it is compared to an empty set of annotations:

$$\text{Par}(\text{Sg1}, \varnothing) = 1 \qquad \textbf{(Prop 3.1)}$$

And conversely:

$$\text{Par}(\varnothing, \text{Sg1}) = 0 \qquad \textbf{(Prop 3.2)}$$

- The particularity of a set Sg1 of annotations compared to a set Sg2 does not depend on the elements of Sg2 that do not belong to Sg1:

$$\text{Sg3} \cap \text{Sg1} = \varnothing \Rightarrow \text{Par(Sg1, Sg2)} = \text{Par(Sg1, Sg2} \cup \text{Sg3)} \quad \textbf{(Prop 4)}$$

5.2.5 Measure of semantic particularity

In order to compute the particularity of Sg1 compared to Sg2, we focus on the terms of Sg1 that are not members of Sg2. This requires to address two problems: the terms are not independent, and they do not convey the same amount of information.

Some of the terms of Sg1 that are not members of Sg2 may be linked in the graph. Taking several linked terms into account would result in considering them several times. For example, in Figure 5.4B, considering both "RNA binding" and "tRNA binding" would result in counting twice the contribution of "RNA binding". Therefore, we should only focus on the terms of Sg1 that do not have any descendant in Sg1 and that are not members of Sg2. Some of these terms might be ancestors of terms of Sg2 and should be considered as common to Sg1 and Sg2. We call Sg* the union of Sg and the sets of ancestors of each element of Sg. We call MPT(Sg1, Sg2) the set of most particular terms of Sg1 compared to Sg2. MPT(Sg1, Sg2) is the set of terms of Sg1 that do not have any descendant in Sg1 and that are not members of Sg2*. In the Figure 5.4B, MPT(hsa, dme) = {"tRNA binding"}. Note that MPT(hsa, dme) is composed of one term and not five.

Using the set theory, we could define Par(Sg1, Sg2) as the proportion of elements of Sg1 that belong to MPT(Sg1, Sg2). When computing card(MPT(Sg1, Sg2)), all the elements have the same weight. However, considering the semantics underlying these elements, some of them may be more informative than others and should ideally be emphasized. Different strategies, similar to those already proposed for the computation of the semantic similarity, can be applied.

We then define PI(Sg1, Sg2), the particular informativeness of a set of GO terms Sg1 compared to another set of GO terms Sg2, as the sum of the differences between the informativeness (I) of each term t_p of MPT(Sg1, Sg2) and the informativeness of the most informative common ancestor (MICA) between t_p and Sg2. The PI of a set of terms is the information that is not shared with the other set.

$$PI(Sg1, Sg2) = \sum_{t_p \in MPT(Sg1, Sg2)} I(t_p) - I(MICA(t_p, Sg2)) \tag{5.1}$$

In the Figure 5.4B, PI(hsa, dme) = I(tRNA binding) - I(binding). We have no sum in this example since MPT(Sg1, Sg2) only contains one term.

We last normalize PI to compute Par(Sg1, Sg2), the semantic particularity of the set of GO terms Sg1 compared to the set of GO terms Sg2. We define MCT(Sg1, Sg2), the set of the most informative common terms of Sg1 and Sg2, as the set of the terms belonging to the intersection of Sg1* and Sg2* that do not have any descendant either in Sg1* or in Sg2*. In the Figure 5.4B, MCT(hsa, dme) = {"protein transporter activity", "protein binding"}. Par(Sg1, Sg2) is the ratio of PI(Sg1, Sg2) and the sum of the informativeness of Sg1 most informative terms (i.e. those Sg1-specific and those common with Sg2; the MICA in the PI formula for the Sg1-specific guarantees that the informativeness of common terms is not counted twice).

$$Par(Sg1, Sg2) = \frac{PI(Sg1, Sg2)}{PI(Sg1, Sg2) + \sum_{t_c \in MCT(Sg1, Sg2)} I(t_c)} \tag{5.2}$$

For the example of the Figure 5.4B, this formula becomes:

$$Par(hsa, dme) = \frac{I(tRNA\ binding) - I(binding)}{(I(tRNA\ binding) - I(binding)) + (I(p.\ trsp.\ activity) + I(protein\ binding))} \tag{5.3}$$

Several measures of informativeness have been proposed. The widely used Information Content (IC) family is based on annotations frequencies determined with an appropriate corpus such as the GOA database. The most frequent terms are considered to be the least informative. When considering Gene Ontology annotations, it is necessary to take the GO subsumption hierarchy into account in order to also consider implicit annotations to the terms ancestors [140]. The alternative approach is corpus-independent. A term informativeness is a function of its distance to the root. It is typically used when a relevant corpus cannot be computed (for comparing elements from several species) or does not exist (for poorly studied species). Wang's Semantic Value (SV) computes this type of informativeness. The relevance of the results obtained by this approach has previously been demonstrated [193, 198].

As shown in the equation 5.3, four terms are involved in the calculation of the MF particularity of the human Exportin-5 ortholog compared to the drosophila Exportin-5 ortholog. This comparison is cross-species, so a semantic value-based informativeness measure is relevant. According to the previous formula, the semantic values of the terms involved in the equation 5.3 are: SV(tRNA binding) = 4.201, SV(binding) = 1.8, SV(protein transporter activity) = 2.952 and SV(protein binding) = 2.44. Consequently, we can compute: Par(hsa, dme) = 0.308. Likewise, for Figure 5.4A, Par(hsa, rno) = 0.082.

5.2.6 Use case: *Homo sapiens* aquaporin-mediated transport

We aimed to study a large dataset in order to determine the frequency and the importance of pairs of similar genes where (at least) one of them also has a high particularity value. We used a dataset composed by 51 well-annotated human genes involved in the aquaporin-mediated transport pathway for *Homo sapiens*. We used the list of all involved genes provided by the Reactome database [237]. We computed the Wang similarity and S-Value-based particularities for each pair of genes of this list. As the Human annotation database is one of the most comprehensive, we also duplicated the study using Lin's measure as an IC-based similarity, and IC as a value of GO term informativeness for our specificity. Tables 5.1, 5.2 and 5.3 present the average, standard deviation, minimum and maximum values of particularity measured in this study for each branch of GO. We classified these statistics in 20 similarity categories containing all the comparison results ranging from sim = 0.5 to sim = 0.999 with steps of sim = 0.025.

As similarity increases, particularity tends to decrease, as expected. In each 20 categories in the human aquaporin-mediated transport pathway, some of the genes have an important particularity compared to the others. This demonstrates that our method combining semantic similarity and particularity identifies genes that cannot be identified using only a similarity measure.

Figure 5.5 on page 115 illustrates this case giving the MF annotation graph of two couples of genes: AQP8 and AQP5 in part A and AQP6 and AQP3 in part B. The corresponding similarity and particularity values are presented in table 5.4 on page 116. Both pairs of genes share the same set of common annotations (in blue), and their respective similarity values were close (0.704 for AQP8 and AQP5; 0.696 for AQP6 and AQP3). As AQP8 has no specific annotation, Par(AQP8, AQP5) = 0. Conversely, AQP5 only has two general specific annotations and Par(AQP5, AQP8) = 0.19. However, AQP6 and AQP3 each has several precise specific annotations:
Par(AQP6, AQP3) = 0.247 and Par(AQP3, AQP6) = 0.415. The two couples have close similarity values regardless the method used but they show a very different particularity profile, with much higher particularities between AQP6 and AQP3 than between AQP8 and AQP5. The two distinct informativeness measures used to compute the particularity led to the same conclusion.

These results confirm that among similar genes, some also have some particular functions,

BP	S-value-based particularity				IC-based particularity			
Similarity	Average	Std dev.	Min	Max	Average	Std dev.	Min	Max
[0.5-0.524]	0.401	0.2	0.013	0.844	0.562	0.223	0	0.904
[0.525-0.549]	0.386	0.174	0	0.794	0.532	0.284	0	0.89
[0.55-0.574]	0.347	0.199	0	0.707	0.497	0.244	0	0.886
[0.575-0.599]	0.352	0.198	0	0.798	0.502	0.241	0	0.895
[0.6-0.624]	0.315	0.203	0	0.671	0.495	0.208	0	0.794
[0.625-0.649]	0.292	0.145	0	0.629	0.437	0.25	0	0.882
[0.65-0.674]	0.299	0.162	0	0.615	0.439	0.258	0	0.876
[0.675-0.699]	0.229	0.15	0	0.529	0.451	0.216	0.039	0.839
[0.7-0.724]	0.228	0.166	0	0.631	0.403	0.239	0	0.859
[0.725-0.749]	0.22	0.145	0	0.501	0.35	0.233	0	0.727
[0.75-0.774]	0.202	0.108	0	0.482	0.403	0.207	0	0.775
[0.775-0.799]	0.178	0.118	0	0.563	0.319	0.222	0	0.671
[0.8-0.824]	0.177	0.106	0	0.418	0.31	0.209	0.043	0.646
[0.825-0.849]	0.125	0.071	0	0.327	0.258	0.184	0	0.589
[0.85-0.874]	0.105	0.131	0	0.418	0.201	0.136	0	0.625
[0.875-0.899]	0.061	0.066	0	0.248	0.179	0.123	0	0.651
[0.9-0.924]	0.039	0.061	0	0.211	0.207	0.156	0	0.614
[0.925-0.949]	0.041	0.067	0	0.248	0.193	0.181	0	0.572
[0.95-0.974]	0.032	0.041	0	0.111	0.099	0.076	0	0.196
[0.975-0.999]	0.005	0.006	0	0.015	0.077	0.152	0	0.519

Table 5.1: Particularity value statistics in 20 similarity values ranges from case 2 – BP measures.

and show that this situation can be observed throughout the full range of similarity values. Therefore, a particularity measure is a relevant complement to a similarity measure in order to identify similar elements that also present some particular trait.

CC Similarity	S-value-based particularity				IC-based particularity			
	Average	Std dev.	Min	Max	Average	Std dev.	Min	Max
[0.5-0.524]	0.353	0.233	0	0.846	0.621	0.244	0	0.911
[0.525-0.549]	0.36	0.214	0	0.819	0.707	0.15	0.185	0.977
[0.55-0.574]	0.33	0.187	0	0.799	0.64	0.202	0	0.897
[0.575-0.599]	0.341	0.185	0	0.752	0.613	0.194	0	0.896
[0.6-0.624]	0.317	0.183	0	0.754	0.621	0.165	0	0.888
[0.625-0.649]	0.268	0.18	0	0.706	0.592	0.207	0	0.852
[0.65-0.674]	0.28	0.177	0	0.656	0.553	0.227	0	0.888
[0.675-0.699]	0.24	0.177	0	0.583	0.495	0.241	0	0.845
[0.7-0.724]	0.13	0.159	0	0.543	0.466	0.24	0	0.825
[0.725-0.749]	0.196	0.151	0	0.579	0.428	0.268	0	0.82
[0.75-0.774]	0.134	0.122	0	0.484	0.383	0.246	0	0.819
[0.775-0.799]	0.15	0.127	0	0.489	0.391	0.267	0	0.768
[0.8-0.824]	0.144	0.093	0	0.269	0.19	0.187	0	0.625
[0.825-0.849]	0.133	0.123	0	0.421	0.352	0.231	0	0.73
[0.85-0.874]	0.146	0.152	0	0.373	0.255	0.216	0	0.624
[0.875-0.899]	0.051	0.051	0	0.11	0.145	0.152	0	0.381
[0.9-0.924]	0.067	0.085	0	0.269	0.095	0.095	0	0.189
[0.925-0.949]	-	-	-	-	-	-	-	-
[0.95-0.974]	-	-	-	-	0.131	0.131	0	0.262
[0.975-0.999]	0.012	0.012	0	0.024	0.049	0.049	0	0.098

Table 5.2: Particularity value statistics in 20 similarity values ranges from case 2 – CC measures.

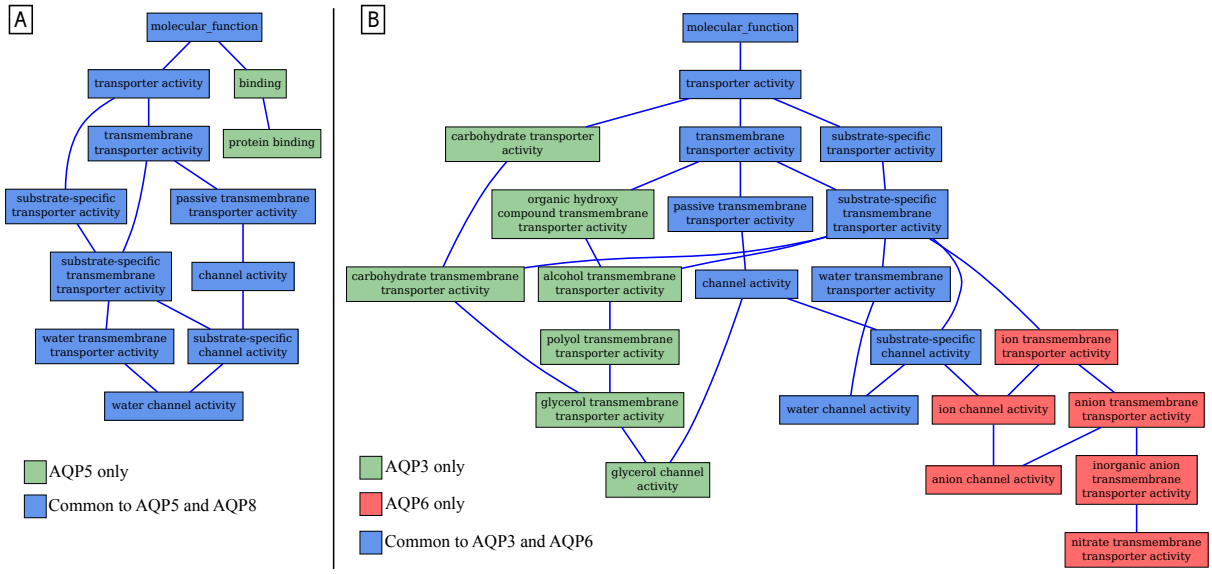

Figure 5.5: MF annotations of two couples of human aquaporins. Part A: AQP8 and AQP5 share most of their annotations. Part B: AQP6 and AQP3 share numerous molecular functions, but each gene also have particular functions. Note that the sets of common annotations are the same in both situation, leading to close similarity values. The respective semantic particularity values reflects AQP3 and AQP6 specific functions, enabling the identification of different patterns.

MF	S-value-based particularity				IC-based particularity			
Similarity	Average	Std dev.	Min	Max	Average	Std dev.	Min	Max
[0.5-0.524]	0.341	0.26	0	0.798	0.494	0.162	0.296	0.701
[0.525-0.549]	0.35	0.219	0	0.818	0.429	0.212	0	0.703
[0.55-0.574]	0.364	0.32	0	0.731	0.422	0.265	0	0.849
[0.575-0.599]	0.382	0.265	0	0.694	0.378	0.148	0.125	0.591
[0.6-0.624]	0.242	0.079	0.132	0.47	0.397	0.205	0	0.81
[0.625-0.649]	0.207	0.113	0	0.531	0.302	0.145	0.158	0.475
[0.65-0.674]	0.281	0.106	0.117	0.482	0.609	0.137	0.13	0.806
[0.675-0.699]	0.223	0.181	0	0.562	0.453	0.249	0	0.763
[0.7-0.724]	0.26	0.267	0	0.564	0.389	0.248	0	0.806
[0.725-0.749]	0.179	0.176	0	0.482	0.419	0.211	0	0.763
[0.75-0.774]	0.171	0.177	0	0.371	0.315	0.216	0	0.643
[0.775-0.799]	0.125	0.167	0	0.482	0.33	0.241	0	0.777
[0.8-0.824]	0.063	0.056	0	0.137	0.239	0.218	0	0.574
[0.825-0.849]	0.119	0.13	0	0.415	0.316	0.222	0	0.574
[0.85-0.874]	0.041	0.036	0	0.116	0.266	0.175	0	0.531
[0.875-0.899]	0.045	0.05	0	0.126	0.179	0.093	0.086	0.272
[0.9-0.924]	0.024	0.025	0	0.055	0.163	0.153	0	0.388
[0.925-0.949]	0.02	0.026	0	0.086	0.09	0.107	0	0.272
[0.95-0.974]	0.005	0.007	0	0.023	-	-	-	-
[0.975-0.999]	-	-	-	-	-	-	-	-

Table 5.3: Particularity value statistics in 20 similarity values ranges from case 2 – MF measures.

SV-based		AQP6	AQP3
Sim	AQP6	1	0.696
	AQP3		1
Par	AQP6	0	0.247
	AQP3	0.415	0

IC-based		AQP6	AQP3
Sim	AQP6	1	0.81
	AQP3		1
Par	AQP6	0	0.531
	AQP3	0.388	0

SV-based		AQP8	AQP5
Sim	AQP8	1	0.704
	AQP5		1
Par	AQP8	0	0
	AQP5	0.19	0

IC-based		AQP8	AQP5
Sim	AQP8	1	0.8
	AQP5		1
Par	AQP8	0	0
	AQP5	0.13	0

Table 5.4: Similarity and particularity values of two couples of genes from case 2. The similarity between AQP6 and AQP3 is very close to the similarity between AQP8 and AQP5 regardless the method used (SV or IC-based). However, the particularity profile obtained for each couple is very different. Again, the SV-based and IC-based methods led to the same conclusion.

5.3 Methodology: threshold determination for similarity and particularity

As we have seen in Figure 5.5 on the previous page, AQP5 and AQP8 are similar, and so are AQP3 and AQP6. However, AQP3 and AQP6 each exhibits some specific function, contrary to AQP5 and AQP8. This interpretation is supported by the numeric values of their respective

semantic semantic similarities and particularities, as shown in table 5.4 on the facing page.

In order to be able to automatize the interpretation of these values, we have to determine the value above which two entities can be considered similar or particular.

This study focuses on a method for determining semantic similarity and particularity thresholds for the interpretation of semantic comparisons. As we have seen in the previous section, this interpretation consists in associating the similarity and particularity values to some similarity and particularity pattern (e.g. two genes are similar and the second gene has a particular function). This section presents the general principle for determining similarity thresholds on the Gene Ontology, and studies the threshold robustness. We then show how this principle is also applicable for determining particularity thresholds on the Gene Ontology. Eventually, we performed an extensive systematic comparison of the thresholds we computed with the traditional 0.5 over the HomoloGene database. This showed that in 5.4% of the comparisons, the thresholds resulted in different patterns. Overall, the new thresholds increased the detection of the "similar with some particularity" pattern, and decreased the number of the inconsistent "similar and both particular" and "neither similar nor particular" patterns. We then focused on the PPAR multigene family and showed that the similarity and particularity patterns obtained with our thresholds discriminated orthologs and paralogs better than those obtained using default thresholds.
In retrospect, this work is interesting because the interpretation of similarity measures usually hinges on implicit thresholds (e.g. "a similarity of 0.83 is *high enough* to consider that two genes are similar") or arbitrary ones (e.g. 0.5 for measures in [0;1]). However, no systematic study had been carried on for determining what these thresholds should be. This study proposes a generic method for determining the optimal threshold for semantic similarity measures and their associated particularity. It is applicable to any ontology and any semantic similarity and particularity measure. In a previous study, we had shown that the ongoing evolution of ontologies such as GO modifies their structure, which in turn can affect the threshold value [118]. Therefore, the thresholds obtained by our method should be regularly updated.

This study was originally published in: Charles Bettembourg, Christian Diot, and Olivier Dameron. Optimal threshold determination for interpreting semantic similarity and particularity: Application to the comparison of gene sets and metabolic pathways using GO and ChEBI. *PloS ONE*, 10(7):e0133579, 2015.

The original article performed the study on the three axis of the Gene Ontology: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). Only BP is detailed here.

The original article also shows that our threshold determination method is applicable to other ontologies such as the Chemical Entities of Biological Interest ontology (ChEBI) [238]. This is not detailed here.

5.3.1 Context

In the previous section, we proposed to combine semantic similarity measures and a new semantic particularity measure to improve the results of gene set analysis [96]. Data analysis often hinges on a qualitative interpretation of the similarity values in order to contrast similar and dissimilar pairs of genes. This discretization of the similarity and particularity values makes the

interpretation easier. It determines whether a functional difference between two genes is or is not marginal.

The main focus of studies to date has been on defining the measures, but **there is no extensive study on the interpretation of the values obtained with these measures**. There has neither been any systematic analysis of the optimal threshold value separating similar from dissimilar. As a result, interpretation is frequently based on either an implicit threshold (for example: "a similarity of 0.83 is *high enough* to consider that two genes are similar" without mentioning when a value reaches this point) or an arbitrary one (typically 0.5 for measures in [0;1] even though no mathematical property of the measures supports this choice).

There are cases where a threshold of 0.5 may be ill-adapted. For example, the similarity value between protein tyrosine kinase 2 (PTK2) and Ubiquitin B (UBB) is 0.502 using Wang's similarity measure on their Biological Processes (BP) annotations. This value is just above the intuitive mid-interval threshold. These two genes are well annotated, with 73 and 79 distinct BP annotations, respectively. According to Entrez Gene, PTK2 is involved in cell growth and intracellular signal transduction pathways triggered in response to certain neural peptides or cell interactions with the extracellular matrix while UBB is required for ATP-dependent, nonlysosomal intracellular protein degradation of abnormal proteins and normal proteins with rapid turnover. These processes cannot be considered "similar". Consequently, the 0.502 value of similarity should not lead to consider PTK2 and UBB as similar genes according to the BP they participate in.

The main factors influencing the similarity values are: granularity differences in GO, GO topology differences between BP, MF and CC, quantity and "quality" of gene annotations, GO temporal evolution [118]. There is a need for a systematic study of semantic measure values in order to determine optimal similarity and particularity thresholds for the qualitative part of functional gene set analysis. Note that the method for determining these thresholds should also be applicable to all semantic similarity categories as well on other ontologies outside GO.

5.3.2 Objective

We propose a generic method to define suitable thresholds based on analysis of the distributions of similarity values. We then extend this method to the semantic particularity measure. We show that our method is applicable to a node-based and a hybrid semantic similarity measure on the Gene Ontology as well as to the corresponding semantic particularity measures. We study the robustness of our method by applying it to multiple sets of genes. We evaluate our method by determining whether the new thresholds lead to different interpretations, and whether these new interpretations are biologically relevant.

5.3.3 Similarity threshold

5.3.3.1 Method for determining similarity thresholds

We first present the general process. We then provide more details about steps two and three.

General process Figure 5.6 on page 120 illustrates the process for determining a similarity threshold. This process is composed of three steps:

1. Define at least two different groups of genes for species of interest. Within a group, the genes should share some common characteristics. Genes from different groups should share as few characteristics as possible.

2. (a) In each group, compute the similarities between each pair of genes (i.e. the intra-group similarities). Gather all the similarity results to obtain an S distribution of similar genes.

   (b) Compute the similarities between each combination of a gene from the first group and a gene from a second group (i.e. the inter-group similarities). Gather all the similarity results to obtain an N distribution of non-similar genes.

3. If the S and N distributions have no overlap between the ranges (min, max), define the threshold τ_{sim} using any value between τ_S (the lowest value of S) and τ_N (the highest value of N). Else, there are some false negatives (FN) and some false positives (FP):

   (a) Compute the proportion of FN in the S distribution for all samples of the similarity threshold between τ_N to τ_S. In this step, consider every value under the similarity threshold as a FN.

   (b) Compute the proportion of FP in the N distribution for all samples of the similarity threshold between τ_N to τ_S. In this step, consider every value above the similarity threshold as a FP.

   (c) For each possible threshold value, sum the FN and FP proportions obtained in steps 3a and 3b. The similarity threshold τ_{sim} is the threshold that minimizes this sum.

Constitution of the S and N distributions We ran a statistical test to determine whether the S and N distributions obtained at step 2 are significantly different. As we cannot consider that the S and N variances are similar, we used an unequal variance t-test (Welch's t-test) which is the recommended test when considering different-sized distributions like S and N. Welch's t-test performs better than Student's t-test when the variances are unequal yet still performs on a par with the Student's t-test when the variances are equal [239]. If the test concludes that the S and N distributions are non significantly different, the process has to be restarted at its first step.

Overlap of the S and N distribution The minimization at step 3c has to be done on FN and FP proportions as the N and S distributions have different sizes.

When comparing the distributions of similar genes (S) to non-similar genes (N), if the minimum value of S is smaller than the maximum value of N, then the S and N distributions overlap and any threshold would lead to FPs or FNs.

Figure 5.7 on page 121 illustrates the case without overlap, where $\min(S) = a$, $\max(N) = b$ and $a > b$. A similarity value greater than a means that the genes compared are similar. A similarity value lower than b means that the genes compared are non-similar. A similarity value between a and b means that the genes compared are nearly similar and thus require expert opinion to interpret the result.

Figure 5.8 on page 121 illustrates the case where the S and N distributions overlap, meaning that there are some FPs (i.e. pairs of genes from N that are non-similar but that have a similarity value greater than a) and FNs (i.e. pairs of genes from S that are similar but have a similarity value lower than b). In this case, a similarity value lower than a means that the genes compared are non-similar. A similarity value greater than b means that the genes compared are similar. Again, expert opinion would be required to interpret the result in this interval. However, in this case, it is possible to determine the threshold value that minimizes both FP and FN.

We established a general framework that defines three thresholds values:

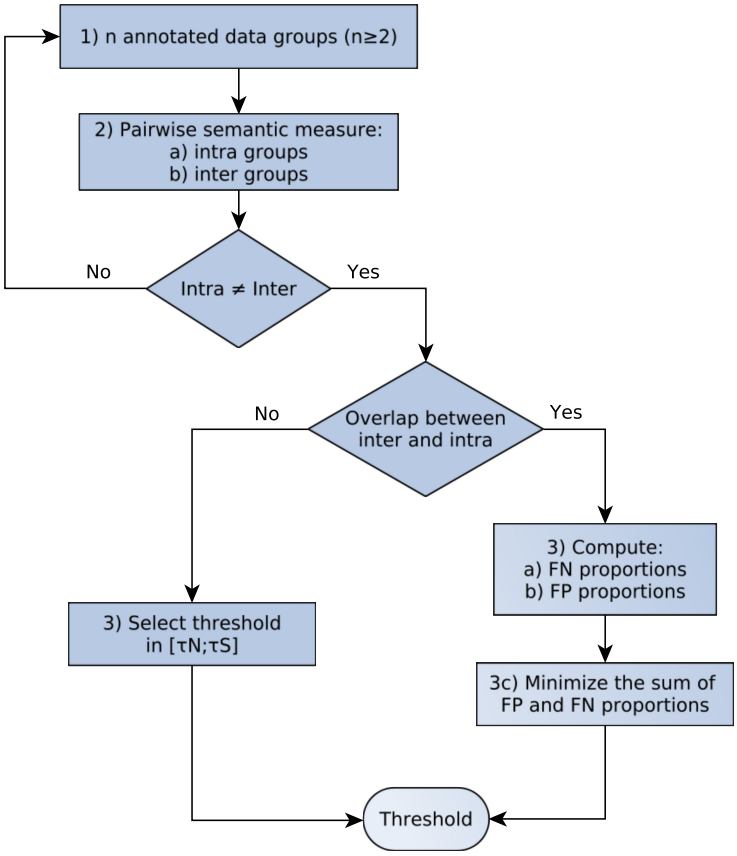

Figure 5.6: Flowchart for threshold determination. 1) Define at least two distinct groups of genes expected to be similar. 2) Compute the intra- and inter-group similarities and compile the results into S and N distributions. If these two distributions are significantly different, the groups of genes are relevant. 3) If S and N do not overlap, define threshold τ_{sim} using any value between τ_S (the lowest value of S) and τ_N (the highest value of N). Else, considering every value under the threshold as FN and every value above the threshold as FP, compute the FN proportion in the S distribution (3a) and the FP proportion in the N distribution (3b) for all samples of the similarity threshold between τ_N to τ_S. 3c) For each possible threshold value, sum the FN and FP proportions obtained in steps 3a and 3b. The similarity threshold τ_{sim} is the one that minimizes this sum.

- $\tau_S = \max(a, b)$ is the threshold value above which the two compared genes are similar. There can not be any FP above τ_S, but there may be some FN below τ_S if $a < b$.
- $\tau_N = \min(a, b)$ is the threshold value under which the two compared genes are non-similar. There cannot be any FN below τ_N, but there may be some FP above τ_N if $a < b$.
- τ_{sim} is the threshold value located between τ_S and τ_N that that minimizes the proportion of FP and FN. As τ_{sim} gets closer to τ_S, there will be more FN and fewer FP. Conversely, as τ_{sim} gets closer to τ_N, there will be more FP and fewer FN. τ_{sim} has to be computed using the proportions of FP and FN as the S and N distributions have different sizes.

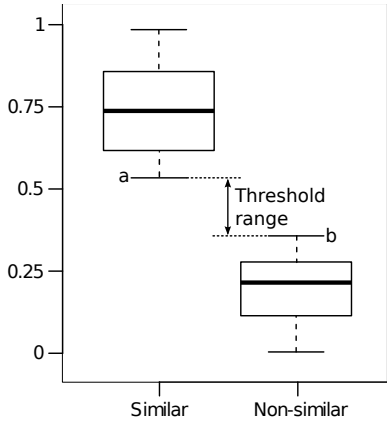

Figure 5.7: Ideal case of threshold determination. The threshold should be located between the lowest whisker of the similar distribution (a) and the upmost whisker of the non-similar distribution (b).

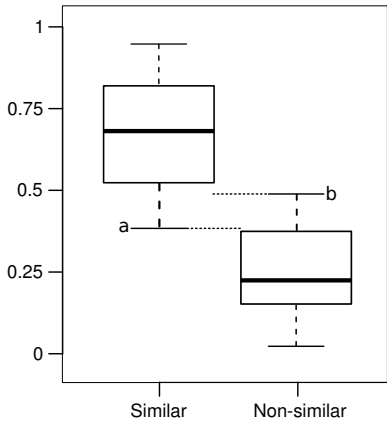

Figure 5.8: Overlap case of threshold determination. The similar and non-similar boxes overlap. In this case, there are false-positive and false-negative results between the lowest whisker of the similar distribution (a) and the upmost whisker of the non-similar distribution (b).

We applied this method to compute Lin's and Wang's semantic similarity thresholds on GO, as well as the corresponding IC-based and SV-based semantic particularity thresholds on GO. For all the pairs of genes compared, we used the GO annotations from the August 2013 version

of GOA. We computed Lin's similarity with the GOSemSim R package [240] (version 1.18.0) using its GO and IC tables and the best-match average approach to compare genes. Pesquita *et al.* showed that the best-match average approach performs best [198]. We computed Wang's similarity, IC-based particularity and SV-based particularity using an in-house implementation of each measure and the August 2013 version of GO.

5.3.3.2 BP similarity threshold using two groups of similar genes

We studied the similarity values obtained when comparing genes known to be functionally close and genes without functional proximity. This study was performed using a hybrid semantic similarity measure (Wang) and a node-based measure (Lin).

Figure 5.9 on the facing page presents the distribution of the BP similarity values obtained for two intra-family comparisons and the corresponding inter-family comparisons. The two PANTHER families were "neurotransmitter gated ion channel" (pthr18945) and "tyrosine-protein kinase receptor" (pthr24416).

As expected, similarity values obtained using either Wang's (Figure 5.9A) or Lin's measure (Figure 5.9B) were significantly higher in the intra-family comparisons than the inter-family comparisons (Welch's t-tests). We observed an overlap between the S and N distributions, which corresponds to the situation shown in Figure 5.8 on the previous page. τ_N was located at the lowest whisker of the intra-family S blue box, *i.e.* 0.096 with Wang's measure and 0.364 with Lin's measure. τ_S was located at the upmost whisker of the inter-family N yellow box, *i.e.* 0.519 with Wang's measure and 0.588 with Lin's measure.

We also determined the optimal similarity threshold value τ_{sim} that minimizes the sum of FP and FN proportions. Figure 5.10 on page 124 reports the results for Wang's and Lin's measures. The minimum ordinate value of the curves gives the threshold for BP using Wang's (0.42) and the Lin's (0.49) measures, respectively.

We used a similar approach for CC and MF; see original article.

5.3.3.3 Robustness of threshold determination

The more groups we build to constitute the S and N distributions, the more reliable the thresholds obtained become. We generalized the above-described process using six groups of similar genes for BP in order to determine τ_S, τ_N and τ_{sim} for Wang's and Lin's measures.

We computed the S distribution gathering the similarity values of each pair of genes inside six different PANTHER families. These families were "histone h1/h5 (pthr11467)", "g-protein coupled receptor" (pthr12011), "neurotransmitter gated ion channel" (pthr18945), "tyrosine-protein kinase receptor" (pthr24416), "phosphatidylinositol kinase" (pthr10048) and "sulfate transporter" (pthr11814). We computed the fifteen distributions corresponding to all the combinations of genes similarity values from two of the previous six families. Each of these distributions is composed of the similarity values between each gene from the first family and each gene from the second family. We combined all these inter-family similarity values into a global N distribution.

In each previous case, the S and N distributions overlapped so defining a threshold in this interval yields some FPs and some FNs. We determined the optimal similarity threshold value that minimizes the sum of FP and FN proportions. Figure 5.11 on page 124 reports the results for Wang's SV-based measure and for Lin's IC-based measure. The minimum ordinate value of each curve gives the threshold for BP, MF and CC using Wang's and Lin's measures, respectively. These similarity thresholds differed according to similarity measure used. They also differed between BP, MF and CC. This can be explained by the different level of complexity between these three branches [118]. It is possible to use one of the three proposed thresholds

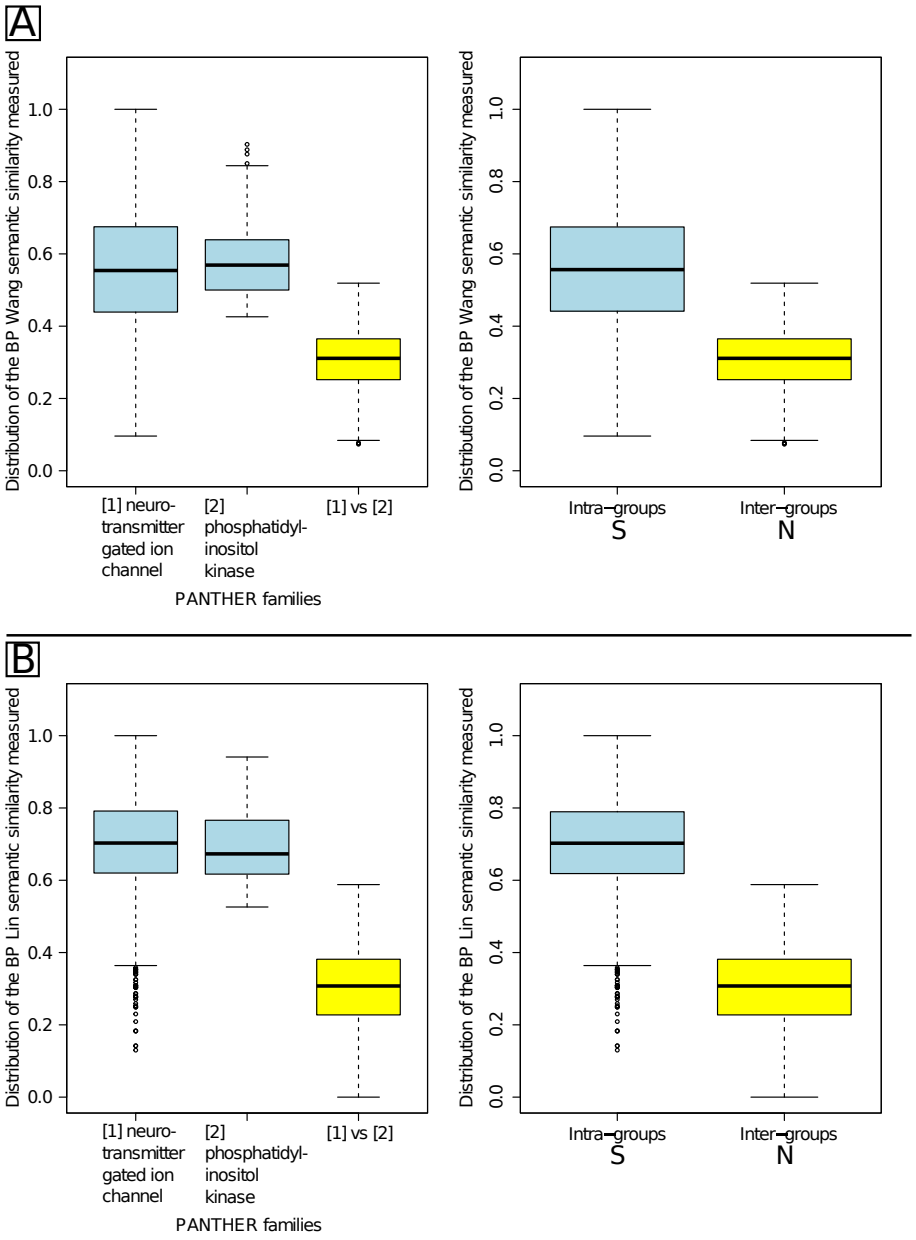

Figure 5.9: Intra- and inter-family semantic similarity distributions using two families of similar genes. Part A presents the results obtained using Wang's measure and part B presents the results obtained using Lin's measure. In both parts, the left side separately presents the two intra-family distributions in blue and the inter-family distribution in yellow. The right side presents the S distribution that gathers all the intra-family similarity values in blue and the N distribution that gathers all the inter-family similarity values in yellow.

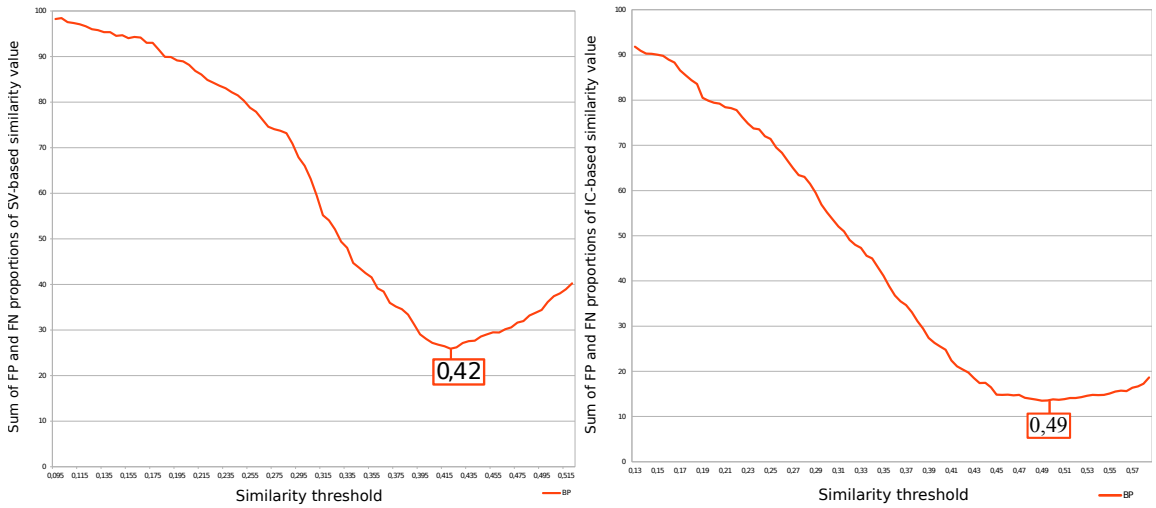

Figure 5.10: Determination of Wang's similarity threshold (left) and Lin's similarity threshold (right) using two families of similar genes. The minimum of false-positive and false-negative proportions gives the similarity threshold (τ_{sim}).

(τ_N, τ_S and τ_{sim}) depending on the accuracy needed to interpret the semantic similarity results. None of these thresholds is equal to the intuitive "default" threshold of 0.5.

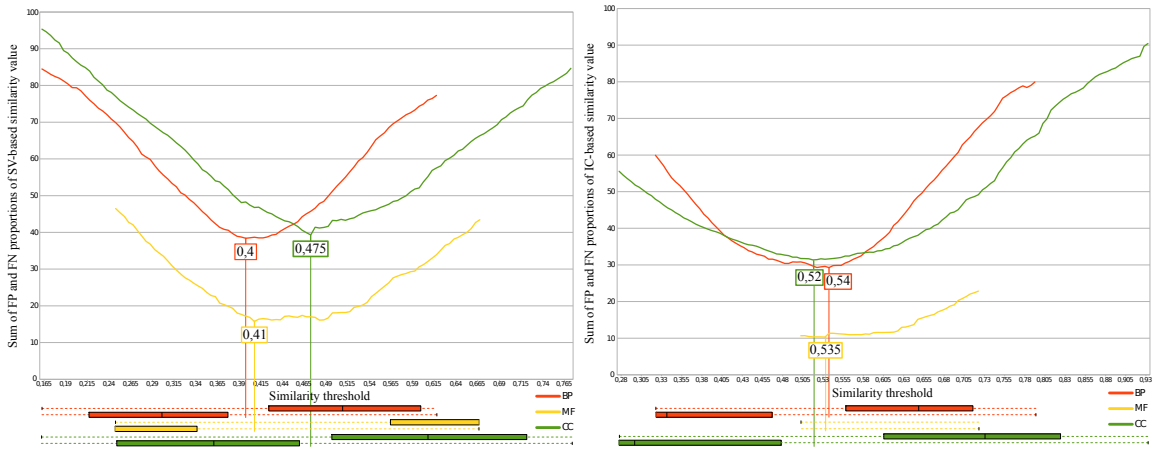

Figure 5.11: Determination of Wang's similarity threshold (left) and Lin's similarity threshold (right). The minimum of false-positive and false-negative proportions gives the similarity threshold (τ_{sim}). The overlapping parts of the boxplots (between τ_N and τ_S) are shown in the lower part of the figure. The thresholds are located between the similar and non-similar boxes.

We validated our study using a leave-one-out approach that consisted in successively recomputing the thresholds using all the sets but one. This approach provides an evaluation of threshold stability.

The thresholds varied slightly over the different datasets. BP similarity threshold varied between 0.4 and 0.435. MF similarity threshold remained stable at 0.41, except when not taking into account the family of genes related to neurotransmitter gated ion channels (0.49). CC similarity threshold was between 0.475 and 0.515.

5.3.4 Particularity threshold

5.3.4.1 Method for determining particularity thresholds

In addition to the similarity thresholds determination, we used the same approach to compute semantic particularity thresholds on BP, CC and MF in order to determine the comparison profile of two genes G1 and G2. The procedure consisted in comparing each value of the triple (Similarity(G1,G2); Particularity(G1,G2); Particularity(G2,G1)) with its respective threshold (noted "+" if the value is greater than the threshold, and "-" otherwise). The results of comparing two genes on their similarity and particularity values can be classified into eight distinct patterns described in Table 5.5. A comparison should not result in a "+ + +" nor a "- - -" pattern. Indeed, a "+ + +" pattern would mean that the two genes compared share enough features to be considered similar yet, at the same time, that each have enough particular features to both be considered particular. Conversely, a "- - -" pattern would mean that the two genes compared are neither similar nor particular.

Notation	sim(A, B)	par(A, B)	par(B, A)
+ + +	$\geqslant \tau_{sim}$	$\geqslant \tau_{par}$	$\geqslant \tau_{par}$
+ + -	$\geqslant \tau_{sim}$	$\geqslant \tau_{par}$	$< \tau_{par}$
+ - +	$\geqslant \tau_{sim}$	$< \tau_{par}$	$\geqslant \tau_{par}$
+ - -	$\geqslant \tau_{sim}$	$< \tau_{par}$	$< \tau_{par}$
- + +	$< \tau_{sim}$	$\geqslant \tau_{par}$	$\geqslant \tau_{par}$
- + -	$< \tau_{sim}$	$\geqslant \tau_{par}$	$< \tau_{par}$
- - +	$< \tau_{sim}$	$< \tau_{par}$	$\geqslant \tau_{par}$
- - -	$< \tau_{sim}$	$< \tau_{par}$	$< \tau_{par}$

Table 5.5: Patterns of similarity and particularity. The results of a semantic comparison of gene annotations can be classed into eight macro-patterns according to similarity and particularity values. The first sign is a "+" if the similarity is greater than or equal to the similarity threshold τ_{sim}, or a "-" otherwise. The two other signs depends on the two particularity values, a "+" for a particularity greater than the particularity threshold τ_{par} or a "-" otherwise.

We applied the threshold determination process described in Figure 5.6 on page 120 to obtain a particularity threshold. For the first step, we composed the same gene groups as those used to compute the similarity threshold. For the second step, we computed all the intra-group and inter-group particularity values between all possible pairs of genes. At the third step, we did not consider any FPs nor FNs as genes belonging to the same group can have some degree of particularity even if they are similar. However, knowing the similarity threshold, we computed the proportion of "+ + +" and "- - -" patterns found in the results while particularity threshold varied. We computed the particularity threshold τ_{par} using the similarity threshold τ_{sim}. For step 3c, we summed the "+ + +" and "- - -" proportions for each possible particularity threshold value. The particularity threshold τ_{par} was the one that minimized this sum.

5.3.4.2 Computation of particularity thresholds

The variation of the "+ + +" and "- - -" profiles in our datasets was studied using the similarity threshold τ_{sim} obtained in the previous section and sampling the value of τ_{par}, the particularity threshold. Table 5.6 on the following page gives the particularity thresholds (τ_{par}) minimizing the sum of "+ + +" and "- - -" patterns for SV-based and IC-based approaches.

These thresholds differed between BP, MF and CC and between approaches (Figure 5.12).

	SV-based particularity threshold	IC-based particularity threshold
BP	0.515	0.68
MF	0.485	0.66
CC	0.335	0.6

Table 5.6: Semantic SV-based and IC-based particularity thresholds. These thresholds minimize the proportions of non-informative "+ + +" or "- - -" patterns according to Table 5.5.

We performed the leave-one-out study in order to assess stability of the particularity threshold by removing one gene set from our datasets and re-computing the particularity threshold. This analysis was performed on BP, MF and CC. The thresholds varied slightly among the different datasets:

- BP particularity threshold was between 0.49 and 0.515 ;
- MF particularity threshold was between 0.35 and 0.485 ;
- CC particularity threshold was between 0.28 and 0.335.

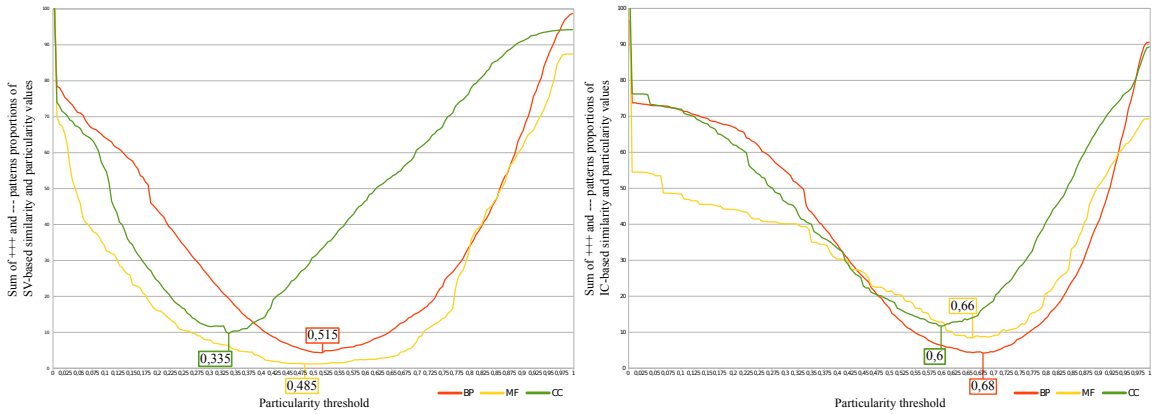

Figure 5.12: Determination of the SV-based particularity threshold (left) and the IC-based particularity threshold (right). The minimum of "+ + +" and "- - -" pattern proportions gives the particularity threshold.

5.3.5 Evaluation of the impact of the new threshold on HolomoGene

The evaluation study involved first quantifying the extent of the changes resulting from using the threshold computed by our method instead of the default 0.5 and then determining whether these changes are biologically relevant.

5.3.5.1 Large-scale evaluation of the impact of threshold changes

We evaluated the impact of our new GO similarity and particularity thresholds over the whole HomoloGene database intra-group gene comparisons. HomoloGene is a system that automatically detects homologs, including paralogs and orthologs, among the genes of 21 fully-sequenced eukaryotic genomes [241].

Table 5.7 summarizes the results for BP. It provides the number of pairs of genes changing from one pattern of Table 5.5 to another using τ_{sim} and τ_{par} instead of the default value of 0.5. We have not distinguished the "+ + -" and "+ - +" categories nor the "- + -" and "- - +" categories as the order of particularity values in the results of this study is meaningless. All categories of the pattern described in Table 5.5 were impacted by the change of threshold. The greatest size increase concerned the "+ + - or + - +" category (+26.2% for BP). The number of "+ + +" and "- - -" cases, that are the least-informative cases, decreased (-11.2% for BP).

BP ↱	+ - -	+ + - or + - +	+ + +	- + +	- + - or - - +	- - -	Total using 0.5 thresholds
+ - -	268,471	0	0	0	0	0	268,471
+ + - or + - +	1,780	54,168	0	0	0	0	55,948
+ + +	7	270	2,623	0	0	0	2,900
- + +	2	154	2,254	10,374	304	1	13,089
- + - or - - +	177	16,027	0	0	32,578	102	48,884
- - -	2,883	0	0	0	0	1,401	4,284
Total using new thresholds	273,320	70,619	4,877	10,374	32,882	1,504	T= 393,576

Table 5.7: **Evolution in patterns in results on HomoloGene intra-group BP comparisons**. Numbers of pairs of genes changing from one pattern to another when considering our optimal similarity and particularity thresholds instead of the default value of 0.5. The most important transition consists in 16,027 results moving from the "- + - or - - +" category (size decreased by 32.7%) to the "+ + - or + - +" category (size increased by 26.2%). The new thresholds give more "+ + +" results but fewer "- - -" results. Globally, the sum of the numbers of the "+ + +" and "- - -" patterns has decreased (-11.2%).

Overall, on BP, CC and MF, the change of thresholds:

- deeply impacted the distribution the HomoloGene intra-group comparison results between the different patterns;
- resulted in important transition from the "- + - or - - +" to the "+ + - or + - +" patterns;
- resulted in fewer "+ + +" and "- - -" cases.

Analysis of relevance on the PPAR multigene family

We measured similarity and particularity values of PPARα, PPARβ and PPARγ between six species. Each gene was only annotated by one or two CC terms, so we kept CC results out of this study. All our similarity values were greater than τ_{sim} (which is not surprising as we are considering genes from the same family). Consequently, in order to emerge similarity differences between orthologs and paralogs, we had to use the more stringent τ_S. This threshold guarantees that the results above it indicate two similar genes. However, the only conclusion that can be inferred for the gene comparisons resulting in values between τ_{sim} and τ_S is that there is doubt over whether these genes are similar. The results of inter-orthologs comparisons systematically matched a "+ - -" pattern, as expected. In contrast, the results of inter-paralog comparisons included some values lower than τ_S and greater than τ_{par}, resulting in "+ + -", "- + -" and "- - +" patterns. Consequently, the thresholds we computed for similarity and particularity measures resulted in patterns consistent with the ortholog conjecture for the PPAR gene family.

5.4 Synthesis

What we learned

- Similarity and particularity metrics allow to provide an objective measure for the comparison of two elements.
- This process can be improved by using ontologies in order to take existing knowledge into account.
- Using semantic particularity as a complement to semantic similarity further refine the analysis when peculiarities are also of biological interest.
- Having a numeric value for similarity and particularity is good because it supports ranking the elements to compare. In our study, we were interested in sorting the pathway steps from the most similar to the least, and among the similar ones, from the most particular to the least.
- Surprisingly, the next step in analysis involved a coarse discretization in order to distinguish the similar elements from the dissimilar ones, and the particular from the non-particular. Although this is done on a regular basis in life science articles, no sound method existed to determine how similar (resp. particular) two elements should be for being considered similar (resp. particular). We proposed an empirical method applicable to any similarity and particularity metrics, over any ontology. This method generated threshold that are different from the usual implicit thresholds, biologically-relevant, rather robust (i.e. choosing a slightly different threshold only has a small impact on the performances) and that can be recomputed when ontologies evolve.

Chapter 6

Conclusion and research perspectives

Since 2003, my research interests have gradually evolved from the representation of symbolic knowledge in semantically-rich formalisms and the associated reasoning to the development of similarity and particularity-based reasoning on semantically simpler ontologies. This transition resulted from both my growing interest in bioinformatics, and from the fortunate conjunction at that time of biological data becoming increasingly available as part of the (open) linked data initiative (which is more difficult in the medical domain), and of the release of SPARQL1.1. Indeed, in bioinformatics answering biologically-relevant questions involves data integration and comparison rather than classification, and, as we have seen, SPARQL1.1 supports most of the needs for querying and integrating data annotated with simple ontologies.

In this context, the reasoning methods I developed gave encouraging preliminary results on several projects Dyliss is currently involved in. However, both our production and usage of linked data is still fragmentary, *ad hoc* and incomplete. It is becoming clear that in each project, we are facing the same limitations.

My research perspectives build on my previous works to tackle the challenges of producing and querying linked data, as well as developing semantic-based methods for analyzing complex life science data.
The first strategic requirement consists in setting up an environment for representing our research data as linked data, ideally with the support of the GenOuest platform. This task encompasses the conversion of data as well as the development of a virtual research environment. Beyond the engineering aspect, the open research challenge lies in the development of a consistent data and metadata management methodology.
The second strategic requirement consists in querying these data. Again, we expect that some data analysis patterns are common to several projects, which requires to store and share them, independently from the datasets. Moreover, we should be able to formulate new relevant biological questions that used to be out of reach when data were more scarce and when combining and processing data was more difficult than it currently is (and will be as we make progress on the first requirement). Eventually, the combination of structurally and semantically-rich data becoming available and of complex queries call for tools capable of abstracting this complexity for the user.
The third strategic requirement focuses on methods for analyzing the data uncovered by the complex queries of the previous requirement. The results of such queries are typically so large and complex that they are themselves useless, until we develop dedicated analysis methods. Again, this is a general problem, so I expect these methods to hinge on a core of generic reasoning primitives that will probably involve domain knowledge for interpretation and will be applicable to multiple projects, at least for metabolic network analysis.

6.1 Producing and querying linked data

Over the last few years, most of the major life science data and knowledge consortia have provided access to some RDF version of their data: pathway databases such as Reactome [237], Wikipathway [242], *CYC [243] are available in the BioPAX format [244, 245, 246] (as they remain incomplete, their integration is desirable but remains a challenge of its own [247, 248, 249]). Others are even providing dedicated SPARQL endpoints that will support federated queries: Uniprot[1] [250], resources from the EBI[2] [251] (currently BioModels, BioSamples, ChEMBL, Expression Atlas and Reactome and others are in preparation) or PubChemRDF [252]. Moreover, initiatives such as identifiers.org[3] simplify the integration of life science data identifiers from different sources [64]. Eventually, repositories such as bio2rdf[4] [61, 63] and BioPortal[5] [49] offer some uniform access to respectively 35 datasets and 442 ontologies.

However, all these are typically resources that we use when analyzing our data, but none our data themselves are currently in RDF. This makes the analysis cumbersome as we have to develop *ad hoc* conversion scripts that hamper exploratory work.

I intend to address the following two challenges that we encounter repeatedly:

- **incorporate the data we work with into the linked data framework** for our direct benefit (analyzing our data better and giving them a better visibility) as well as for the

[1] `http://sparql.uniprot.org/`
[2] `https://www.ebi.ac.uk/rdf/platform`
[3] `http://identifiers.org/`
[4] `https://github.com/bio2rdf`
[5] `http://bioportal.bioontology.org/`

benefit of the community [8]

- when (linked) data are here, we still have to **invent the querying that takes full advantage of the linked data framework**. This encompasses two problems: (1) a bioinformatics one about formulating new biological questions that are relevant but used to be out of reach for lack of available data and querying capabilities [253], and (2) a computer science one about providing an infrastructure for representing these data and for supporting their new querying (which will most probably involve the Semantic Web). I expect to focus on the first one but will rely extensively rely on the second one, with some possible marginal contributions.

6.1.1 Representing our data as linked data

Incorporating our data in the linked data framework consists in storing and sharing our data as well as linking them to other resources such as genes, pathways, RNA fragments, taxons, molecules or proteins. While data storage will consist in using available technical solutions such as Virtuoso[6] or Fuseki[7], making explicit the relations to other resources and integrating everything into an E-Science context that remains to be developed goes beyond simple engineering.

As we see below, this challenge is common to many projects.

6.1.1.1 Converting data into RDF

MiRNAdapt on aphids The MiRNAdapt project[8] led by Denis Tagu from INRA aims at studying how aphids' gene expression adapt to changes of the local environment such as seasons, and particularly genic regulation during pea aphid embryogenesis. The project produced large quantities of data about messenger RNA, microRNA, piRNA and long non-coding RNA expression levels as well as epigenetic marks such as histone and DNA methylation. These data are stored in 16 tabulated flat files totaling 6,160,765 lines. Analyzing these data requires to be able to query them uniformly even if they were obtained separately, as well as connecting then with external resources [254]. Currently, the biologists import the files as spreadsheets for being able to process them. The processing must be manually adapted and repeated for each new query, which usually takes between two and three hours each time before computation can take place.

Since 2014, with Fabrice Legeai, Anthony Bretaudeau and Charles Bettembourg, we imported the information in RDF (45,278,179 triples) and stored it in a triplestore [101]. This process only needs to be done once. We were then able to write SPARQL queries for each 6 use cases, which demonstrated that SPARQL has the necessary expressivity. Writing each query took only a few minutes, and the queries can be reused and adapted, which simplifies analysis, particularly exploratory hypotheses. For this proof of concept, the flat file conversion in RDF was performed with *ad hoc* scripts. We are investigating how to streamline this process, e.g. using tarql[9] or directly in the triplestore[10].

EPICLUB on *Brassicaceae* The EPICLUB project led by Mélanie Jubault from INRA aims to determine the respective parts of epigenetics and genetics in *Brassicaceae* (cabbages, broccoli, cauliflower, Brussels sprouts, radishes,...) response to clubroot, a common disease

[6] `http://virtuoso.openlinksw.com/`

[7] `https://jena.apache.org/documentation/fuseki2/`

[8] `http://www6.rennes.inra.fr/igepp_eng/RESEARCH-TEAMS/Ecology-and-Genetics-of-Insects/Projects/MiRNAdapt2`

[9] `http://tarql.github.io/`

[10] `http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtCsvFileBulkLoader`

caused by a protist called *Plasmodiophora brassicae*. The project requires an infrastructure for managing and integrating a large quantity of data including *Brassicaceae* genome sequences with their orthology and synteny relations, resistance major genes and QTL, and transcriptomics, metabolomics and epigenomics data. Currently, the data are available as text and csv files as well as spreadsheets. With Aurélie Évrard and Mélanie Jubault, we will follow an approach similar to the MiRNAdapt project.

Patient care trajectories The PEPS platform (*plateforme pharmaco-épidémiologie des produits de santé*) led by Emmanuel Oger (CHU Rennes) aims at providing an infrastructure for performing large scale pharmacoepidemiology studies based on French national medico-administrative databases such as SNIIRAM (*Système National d'Information Inter-Régime de l'Assurance Maladie*, the French equivalent to National Health Insurance Cross-Schemes Information System NHI-CIS) for healthcare reimbursement (e.g. drug prescriptions, medical transports) and PMSI (*Programme de Médicalisation des Systèmes d'Information*) from hospital discharge information systems (e.g. diagnosis and procedures). In collaboration with Nolwenn Le Meur and Yann Rivault (EHESP), we focused on detecting complications for patients having an day surgery (i.e. a surgery that does not require an overnight hospital stay) between January and December 2012, and their follow-up data over 2013. The dataset concerned 1,389,271 patients and 1,636,445 instances of procedure. We wanted to determine which patients had an outpatient surgery, and among them which ones had a pattern suggesting a possible complication (for example an antibiotics prescription or another hospitalization in the following days). The data were too big for being handled by R (which was not a surprise), and our need to use ontologies about procedures, drugs or diseases made the use of a relational database impractical.

We converted the data in RDF and linked with ontologies such as ATC for drugs, CCAM for procedures and ICD10 for diagnosis. We wrote SPARQL queries to retrieve the patients of interest and their related information [105]. We used R to perform the statistical analysis in order to identify determinants of complications. As in the two previous projects, this turned out to be a relevant solution that supported the need for data integration and data analysis on large datasets. Further work will continue on data representation and on data analysis with the beginning of Yann Rivault's PhD thesis that will focus on the analysis of patients' care trajectories. This will expand a previous work with Gautier Defossez and Alexandre Rollet on the temporal representation of care trajectories of breast cancer patients using data from a regional information system [255], and will benefit from a collaboration with Thomas Guyet, David Gross-Amblard and Yann Dauxais (IRISA).

Synthesis These projects require some graph querying and traversal capabilities, as well as some integration with other resources, for which RDF is well adapted [22]. Some of the analysis methods in use or in development also require some graph topology functions such as finding the maximal cliques, or involve Answer Set Programming, for which RDF may not be the optimal data representation format. Determining whether the data should be stored natively in RDF and exported to other formalisms, or the other way around remains to be investigated for the definition of Dyliss data management plan.

With the help of the GenOuest[11] engineers, these data should be deployed on the GenOuest platform.

[11] `http://www.genouest.org/`

6.1.1.2 Incorporating linked (meta)data into a Virtual Research Environment

All the efforts presented in the previous section are not specific to our team. As we have seen in sections 1.1 and 1.2, life sciences is one of the many domains concerned with the data deluge. Researchers are generating increasing quantities of data in data silos and we are all trying as hard as we can to make things even worse by interconnecting these silos. In translational research, accessing and combining data is a challenge as important as the biological questions we try to answer. We have seen throughout this manuscript that the Semantic Web is valuable for automatically *processing* the data in order to answer biological questions [50]. All the efforts for *managing* the data of the projects presented in the previous sections also suggest that **manual data management specific to each project will fail globally because of both the quantity and the complexity. Tending the scientific information ecosystem should be done systematically**.

E-Science is "both the pursuit of global, collaborative *in silico* science and the computational infra-structure to support it." [256]. In this context, systematic data management relies on metadata associated to the raw data as well as the data produced along the processing steps of the analysis. This is typically supported by Virtual Research Environments (VRE). In bioinformatics, the data processing steps are usually handled by a workflow engine associated with the VRE, such as the Taverna [257] engine with myExperiment [258], or more recently Galaxy[12] [259] and its data manager [260] with HubZero[13].

With the VRE providing an integrated framework for storing the data and the associated metadata, as well as workflow descriptions that can be executed on the data, the next challenge lies in metadata generation. Currently, these metadata are optional: they are not required by any step of the analysis and except for the most simple ones such as the date or the creator, it is up to the user to provide them to the VRE for meeting traceability requirements of for an easier data retrieval. However, this will never scale-up for handling large quantities of data.

With Yvan Le Bras (plateforme GenOuest), Alban Gaignard (institut du thorax Nantes), Audrey Bihouée (plateforme bioinformatique BiRD Nantes), François Moreews (INRA Rennes) and Olivier Collin (plateforme GenOuest), we are working on integrating a semantically-rich metadata (typically based on PROV[14], ISA [261] and EDAM [262]) generation capability into the VRE data management. We hypothesize that the metadata associated with the result of the execution of a workflow can be automatically determined from the annotations of the input data and of the workflow and we propose to embed this capability into the workflow engine itself [263].

Among the strategies we are considering, I propose to:

- create a generic "semantic metadata wrapper" service taking as parameters (1) the identifier of a "regular service", (2) a semantic description of this regular service (the service identifier can be a part of the semantic description) and of its parameters and (3) a semantic description of the workflow invoking the regular service. The semantic wrapper service is responsible for invoking the regular service, and for generating the metadata associated with the result.

- create a service converting a "regular" workflow into a "semantic metadata-enabled" workflow by embedding each service of the original workflow into its semantic metadata wrapper counterpart.

This solution is compatible with any workflow engine, and does not require to hack into the internals of the engine. It requires a minimal amount of manual annotation: once for each

[12] `https://galaxyproject.org/`
[13] `https://hubzero.org/`
[14] `http://www.w3.org/TR/prov-o/`

service and once for each workflow. Moreover, a user can use the regular version of the services when creating a new workflow and then generate the semantically-enabled version for production purpose once he is satisfied. Eventually, the generic wrapper that can possibly be refined using an approach similar to inheritance by creating as many specific wrappers generating special annotations and calling the generic wrapper for handling the general annotations.

6.1.2 Querying linked data

In all the projects mentioned in Section 6.1.1.1, writing SPARQL queries greatly simplifies the analysis. However, in spite of the benefits, this in turn suffers from two main limitations: (1) writing these queries requires a mental representation of the data underlying structure, i.e. what kinds of entities are present and what are the typical relations between them, and (2) not all end-users are willing to take up learning SPARQL, and find it all the more difficult to do so because they also lack (1).

Representations of the structure of the data available at SPARQL end-points are typically provided by additional diagrams (e.g. for Uniprot[15], Reactome[16] or ChEMBL[17]). However, these diagrams are hand-crafted and not always available, which makes the manual exploration of an endpoint cumbersome.

When a diagram is available, the user still has to write SPARQL queries, which (s)he may not be familiar with. This typically consists in writing the SPARQL code in the text area of a website. The most user-friendly solutions feature syntax-highlighting but are still regarded as "too technical", even if some typical example queries and templates are provided. Initiatives such as Sparklis[18] aim at making exploration easier [264]. Sparklis allows the user to build a query step by step by iteratively selecting the relation and the neighbor of a node of interest It was still perceived as lacking ergonomy. Moreover, because of performance constraints, infrequent properties and neighbors may not be presented for a node of interest, which may give the misleading impression that some information is not present in the data.

I propose an unified solution to both problems based on the representation of the data present on a triplestore as a graph, and on a query-building principle using paths on the abstraction graph. Of course this solution should be generic, and will be applicable (among others) to all the projects mentioned in Section 6.1.1.1.

6.1.2.1 RDFmap: building an abstraction graph of data

RDFmap aims at building automatically a graph-based abstraction of a dataset that would be similar to the hand-crafted diagrams used currently.

The general principle consists in identifying the main classes, and in creating a link between two classes if an instance of the first class is associated to an instance of the second class. The whole process can be performed as a SPARQL query. The first results are encouraging. Some work is still required for improving identification of the main classes and for determining which relations between them should be represented.

6.1.2.2 AskOmics: building SPARQL queries as paths on the abstraction graph

AskOmics is being developed by Charles Bettembourg and Fabrice Legeai as a contribution to the MiRNAdapt project [101], but should be applicable to any RDF dataset.

[15] `http://sparql.uniprot.org/images/diagrams/uniprot.jpg`
[16] `https://www.ebi.ac.uk/rdf/sites/ebi.ac.uk.rdf/files/documents/reactome_simplified.png`
[17] `https://www.ebi.ac.uk/chembl/extra/RDF/chembl_18_rdf_summary.png`
[18] `http://www.irisa.fr/LIS/ferre/sparklis/`

It is based on an abstract representation of the MiRNAdapt data which was created manually but should be replaced by the result of RDFmap in the future. AskOmics uses D3.js[19] to provide a visual representation of the abstraction as a graph. By starting from a node of interest and iteratively selecting its neighbors, the user creates a path on the abstraction graph. This path can then be transformed into a SPARQL query that can be executed on the original dataset.

This approach presents several benefits. First, the abstraction graph only needs to be generated once and does not need to be computed on-the-fly for each node, contrary to Sparklis. Second, the whole query building only takes place on the abstraction graph, which is much smaller than a typical RDF dataset. Third, intermediate count() queries can be executed on the dataset to provide on-the-fly hints to the user (or for debugging assistance). Fourth, the same principle can be applied on an optional (manually-generated) "interface layer" on top the abstraction graph in order to hide parts of the abstraction that may not be relevant to the user or to provide "shortcuts" in order to avoid property paths (i.e. composition of relations) ; there can be different interface layers depending on the types of users.

6.2 Analyzing data

I joined the Dyliss team at IRISA in 2013. The team focuses on bioinformatics and systems biology. The main goal in biology is to characterize groups of genetic actors that control the phenotypic answer of non-model species when challenged by their environment. Unlike model species, only a limited prior-knowledge is available for these organisms [30] together with a small range of experimental studies (culture conditions, genetic transformations). To accommodate these limitations, the team explores methods in the field of formal systems, more precisely in knowledge representation, constraints programming, multi-scale analysis of dynamical systems, and machine learning. Our goal is to take into account both the information on physiological responses of the studied species under various constraints and the genetic information from their long-distant cousins.

The challenge to face is thus incompleteness: the limited range of physiological or genetic known perturbations is combined with an incomplete knowledge of living mechanisms involved. We favor the construction and study of a "space of feasible models or hypotheses" including known constraints and facts on a living system rather than searching for a single optimized model. We develop methods allowing a precise investigation of this space of hypotheses. Therefore, the biologist will be in position of developing experimental strategies to progressively shrink the space of hypotheses and gain in the understanding of the system. This refinement approach is particularly suited to non-model organisms, which have specific and little known survival mechanisms. It is also required in the framework of an increasing automation of experimentations in biology.

By exploring the complete space of models, our approach typically produces numerous candidate models compatible with the observations. **My contribution consists in investigating to what extent domain knowledge can further refine the analysis of the set of models by identifying classes of similar models, or by selecting the models that best fit biological knowledge.** We anticipate that this will be particularly relevant when studying non-model species for which little is known but valuable information from other species can be transposed or adapted.

Sections 6.2.1 and 6.2.2 present ongoing works on the selection of relevant candidates when reconstructing metabolic pathways and on the analysis of TGF-β signaling pathways. Although the application domains are different, the reasoning method is strikingly similar in both cases.

[19] `http://d3js.org/`

Section 6.2.3 presents my medium to long range main research goal, that consists in defining a more generic analysis framework combining topological and semantics information.

6.2.1 Selecting relevant candidates when reconstructing metabolic pathways

This work is a contribution to the Idealg project[20] (*investissement d'avenir*). It is a collaboration with Sylvain Prigent, Anne Siegel and Pierre Vignet. The idealg project aims at having a better overall understanding of the three groups of macroalgae (green, red and brown) in order to develop the algae sector in Brittany. This includes especially the study of species specific to each of the three major groups of algae, *Ectocarpus siliculosus* in the case of brown algae.

A part of this projects consists in proposing a complete metabolic network for *Ectocarpus siliculosus*. A metabolic network is the complete set of physical and physiological reactions that explain the overall functioning of a cell. This metabolic network has to have a good quality and has to be compatible with biological observations. It must especially be able to explain the presence of 56 compounds of interest for biologists. Reconstructing metabolic networks is a labor-intensive task requiring numerous biological experiments. Most current efforts relied massively on experts manual intervention either in plants [265, 266, 267] or in animals [268].

Ectocarpus siliculosus not being a "model species", numerous portions of metabolic pathways are unknown [30]. The traditional approach is not applicable in this context because the data are too scarce, would take too long to produce and lack a large-enough community to validate. A classic strategy consists in completing the pathways using reactions observed in other species. However, there are many reactions from many species, spread in several complementary databases [247]. Determining the best candidates from a biological point of view requires incorporating prior knowledge [269] but remains an open challenge, specially for large scale networks [270, 271, 272, 273]. Existing symbolic knowledge represented in ontologies can contribute to address the problem of missing information [274, 275] and the problem of processing large quantities of interdependent data [45]. A previous study used MetaCyc [276] as a source of candidate reactions to complete the metabolic network [99]. The smallest set of MetaCyc reactions to be added to the reconstructed network in order to produce the 56 target proteins of interest is composed of 42 reactions. However, a systematic exploration produced 2400 possible minimal sets that are all structurally equivalent. Together, these 2400 candidate sets cover 70 reactions, so they have a large overlap.

We have developed a knowledge-based method that reduces the number of candidate sets from 2400 to 48. It consists in creating a graph of mutually-exclusive reactions (i.e. couples of reactions that do not belong to any candidate metabolic network) in order to retrieve the maximal cliques. Composing a candidate network requires to select one of the reactions for each clique. We then developed a reasoning method based on ontologies in order to determine for each clique a subset of the reactions that fit best with biological knowledge. Eventually, we only select the candidate networks composed of these reactions. We are currently performing a formal evaluation of this strategy on *Escherichia coli* by artificially degrading metabolic networks before reconstructing them (i.e. to assess whether we selected the relevant candidates and discarded the not-so relevant ones) and investigating further enhancements.

6.2.2 Analyzing TGF-β signaling pathways

This project is a collaboration with Nathalie Théret with whom I supervise Jean Coquet's PhD thesis, Geoffroy Andrieux, Anne Siegel and Jacques Nicolas. The transforming growth factor beta 1 (TGF-β1) protein plays a major role in immune response and in tumor development, as an

[20] `http://www.idealg.ueb.eu/`

antagonist in the early stages and as a promoter in the advanced stages [277]. TGF-β pleiotropic effects are linked to the complex mechanisms regulating its activity. These mechanisms are therefore potential therapeutic targets.

Geoffroy Andrieux developed the most exhaustive discrete model of TGF-β signaling pathways. It contains 9,248 reactions composed of 9,177 components. This model allowed him to identify 15,934 sets of influence composed of chemical reactions and activating some of the 145 genes influenced by TGF-β [278]. The size and the internal complexity of this network prevent its exploitation by biologists.

We are conducting a systematic analysis in order to identify:

- sets of genes activated by similar sets of influence. The relevance of these gene sets will depend on the biological processes of the diseases associated with the genes.
- families of similar sets of influence (i.e. activating the same genes or genes involved in similar processes).
- genes or sets of influence common to several sets and playing the role of interface. Such elements are of potential interest for understanding the transition of TGF-β role from tumor antagonist to tumor promoter.

This analysis consists in a systematic search of associations, and also relies on external domain knowledge such as biological processes or diseases. This knowledge is used both in the search of association and on the interpretation of results.

The analysis consists in determining cliques of genes activated by the same sets of influence, and cliques of sets of influence activating the same genes. We then determine the cliques homogeneity according to biological processes or diseases, and select the most interesting for further analysis by biologists.

6.2.3 Data analysis method combining ontologies and formal concept analysis

The method for selecting candidates after metabolic pathways reconstruction (section 6.2.1) and the method for analyzing signaling pathways (section 6.2.2) both consist of a topological analysis of a domain-dependent graph, followed by a semantic-based method for grouping solutions of for reducing their number. I intend to develop a refined and unified analysis method. The Confocal project (PEPS CNRS FaSciDo 2015) with Anne Siegel, Jacques Nicolas et Nathalie Théret, Jean Coquet, Amedeo Napoli (LORIA Nancy) and Élisabeth Rémy (Institut de Mathématiques de Luminy) is a first step in this direction.

In the biomedical domain, the classical approaches for analyzing annotated elements rely on domain knowledge and semantic similarity values in oder to perform hierarchical clustering [279, 280, 22]. Because data are noisy and incomplete, as mentioned previously, special approaches have been developed [281, 282].

Limitation 1: classical biclustering methods do not permit partial overlap of clusters, which is not compatible with the pleiotropic nature of some genes.

Formal concept analysis (FCA) performs an exhaustive search of maximal sets of elements sharing the same attributes [283]. It addresses the previous limitation and is a relevant alternative because the lattice represents several levels of precision from numerous small sets of genes having many influence sets in common, to fewer larger gene sets sharing fewer influence sets. Contrary to biclustering, the lattice also supports the identification of partially overlapping clusters. Compared to the maximal cliques, it allows us to perform a finer-grain analysis. FCA has already been successfully applied to the analysis of gene expression data [284] and to signaling

networks modeling [285]. Eren et al. have shown that clustering algorithms capable of finding more than one model are more likely to find biologically-relevant clusters [279]. However, FCA also suffers from the following limitations, even if recent breakthrough in the Orpailleur team concerned the combination of biclustering and FCA [286, 287] and a concept stability measure for identifying relevant concepts [288, 289].

Limitation 2: FCA assumes that the elements are independent, whereas we would like to take relations extracted from ontologies into account.

Limitation 3: FCA's exhaustive search generates numerous formal concepts, not all of them being informative (specially the large and the small ones) or biologically-relevant [290].

Limitation 4: FCA is sensitive to noisy and incomplete data. Pensa and Boulicaut developed a fault-tolerant technique that they applied to gene expression analysis [291].

I will focus on using ontologies to guide formal concept analysis for identifying relevant associations among data. This will involve using ontologies (1) before FCA for enriching annotations, and (2) after FCA for identifying semantically-homogeneous clusters that either match existing knowledge (for validation purpose), or do not (for discovery purpose).

Bibliography

[1] Carol J. Bult. From information to understanding: the role of model organism databases in comparative and functional genomics. *Animal Genetics*, 37(suppl. 1):28–40, 2006.

[2] Olivier Bodenreider and Robert Stevens. Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274, 2006.

[3] 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.

[4] Nicola Cannata, Emanuela Merelli, and Russ B. Altman. Time to organize the bioinformatics resourceome. *PLoS Computational Biology*, 1(7):0531–0533, 2005.

[5] R Bellazzi, M Diomidous, I N Sarkar, K Takabayashi, A Ziegler, and A T McCray. Data analysis and data mining: current issues in biomedical informatics. *Methods of information in medicine*, 50(6):536–544, 2011.

[6] Damian Smedley, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, Mohammad Hamza Awedh, Richard Baldock, Giulia Barbiera, Philippe Bardou, Tim Beck, Andrew Blake, Merideth Bonierbale, Anthony J Brookes, Gabriele Bucci, Iwan Buetti, Sarah Burge, Cedric Cabau, Joseph W Carlson, Claude Chelala, Charalambos Chrysostomou, Davide Cittaro, Olivier Collin, Raul Cordova, Rosalind J Cutts, Erik Dassi, Alex Di Genova, Anis Djari, Anthony Esposito, Heather Estrella, Eduardo Eyras, Julio Fernandez-Banet, Simon Forbes, Robert C Free, Takatomo Fujisawa, Emanuela Gadaleta, Jose M Garcia-Manteiga, David Goodstein, Kristian Gray, Jose Afonso Guerra-Assuncao, Bernard Haggarty, Dong-Jin Han, Byung Woo Han, Todd Harris, Jayson Harshbarger, Robert K Hastings, Richard D Hayes, Claire Hoede, Shen Hu, Zhi-Liang Hu, Lucie Hutchins, Zhengyan Kan, Hideya Kawaji, Aminah Keliet, Arnaud Kerhornou, Sunghoon Kim, Rhoda Kinsella, Christophe Klopp, Lei Kong, Daniel Lawson, Dejan Lazarevic, Ji-Hyun Lee, Thomas Letellier, Chuan-Yun Li, Pietro Lio, Chu-Jun Liu, Jie Luo, Alejandro Maass, Jerome Mariette, Thomas Maurel, Stefania Merella, Azza Mostafa Mohamed, Francois Moreews, Ibounyamine Nabihoudine, Nelson Ndegwa, Celine Noirot, Cristian Perez-Llamas, Michael Primig, Alessandro Quattrone, Hadi Quesneville, Davide Rambaldi, James Reecy, Michela Riba, Steven Rosanoff, Amna Ali Saddiq, Elisa Salas, Olivier Sallou, Rebecca Shepherd, Reinhard Simon, Linda Sperling, William Spooner, Daniel M Staines, Delphine Steinbach, Kevin Stone, Elia Stupka, Jon W Teague, Abu Z Dayem Ullah, Jun Wang, Doreen Ware, Marie Wong-Erasmus, Ken Youens-Clark, Amonida Zadissa, Shi-Jian Zhang, and Arek Kasprzyk. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research*, 43(W1):W589–W598, 2015.

[7] Michael Y Galperin, Daniel J Rigden, and Xosé M Fernández-Suárez. The 2015 nucleic acids research database issue and molecular biology database collection. *Nucleic acids research*, 43(Database issue):D1–D5, 2015.

[8] Alyssa Goodman, Alberto Pepe, Alexander W Blocker, Christine L Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, David W Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, and Aleksandra Slavkovic. Ten simple rules for the care and feeding of scientific data. *PLoS computational biology*, 10(4):e1003542, 2014.

[9] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: Astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015.

[10] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[11] Marvalee H Wake. What is "integrative biology"? *Integrative and comparative biology*, 43(2):239–241, 2003.

[12] Jennifer R Tisoncik and Michael G Katze. What is systems biology? *Future microbiology*, 5(2):139–141, 2010.

[13] Bas Teusink, Hans V Westerhoff, and Frank J Bruggeman. Comparative systems biology: from bacteria to man. *Wiley interdisciplinary reviews. Systems biology and medicine*, 2(5):518–532, 2010.

[14] Huajun Chen, Tong Yu, and Jake Y Chen. Semantic web meets integrative biology: a survey. *Briefings in bioinformatics*, 14(1):109–125, 2012.

[15] Francesco M Marincola. Translational medicine: A two-way road. *Journal of translational medicine*, 1(1):1, 2003.

[16] Indra Neil Sarkar. Biomedical informatics and translational medicine. *Journal of translational medicine*, 8:22, 2010.

[17] Atul J Butte. Translational bioinformatics: coming of age. *Journal of the American Medical Informatics Association : JAMIA*, 15(6):709–714, 2008.

[18] Nigam H Shah, Clement Jonquet, Annie P Chiang, Atul J Butte, Rong Chen, and Mark A Musen. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC bioinformatics*, 10 Suppl 2:S1, 2009.

[19] Qing Yan. Translational bioinformatics and systems biology approaches for personalized medicine. *Methods in molecular biology (Clifton, N.J.)*, 662:167–178, 2010.

[20] R B Altman. Translational bioinformatics: Linking the molecular world to the clinical world. *Clinical pharmacology and therapeutics*, 2012. In press.

[21] Atul J Butte and Lucila Ohno-Machado. Making it personal: translational bioinformatics. *Journal of the American Medical Informatics Association : JAMIA*, 20(4):595–596, 2013.

[22] Atsushi Fukushima, Shigehiko Kanaya, and Kozo Nishida. Integrated network analysis and effective tools in plant systems biology. *Frontiers in plant science*, 5:598, 2014.

[23] Yonqing Zhang, Supriyo De, John R Garner, Kirstin Smith, S Alex Wang, and Kevin G Becker. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC medical genomics*, 3:1, 2010.

[24] Kyoohyoung Rho, Bumjin Kim, Youngjun Jang, Sanghyun Lee, Taejeong Bae, Jihae Seo, Chaehwa Seo, Jihyun Lee, Hyunjung Kang, Ungsik Yu, Sunghoon Kim, Sanghyuk Lee, and Wan Kyu Kim. GARNET - gene set analysis with exploration of annotation relations. *BMC bioinformatics*, 12 Suppl 1:S25, 2011.

[25] William A Baumgartner, K Bretonnel Cohen, Lynne M Fox, George Acquaah-Mensah, and Lawrence Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics (Oxford, England)*, 23(13):i41–i48, 2007.

[26] Carole Goble and Robert Stevens. State of the nation in data integration for bioinformatics. *Journal of biomedical informatics*, 41(5):687–693, 2008.

[27] Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.

[28] Hector J. Levesque. On our best behaviour. In *Proceedings of IJCAI2013 conference*, 2013.

[29] Robert Stevens, Carole A. Goble, and Sean Bechhofer. Ontology-based knowledge representation for bioinformatics. *Briefings in bioinformatics*, 1(4):398–416, 2000.

[30] C R Primmer, S Papakostas, E H Leder, M J Davis, and M A Ragan. Annotated genes and nonannotated genomes: cross-species use of gene ontology in ecology and evolution research. *Molecular ecology*, 22(12):3216–3241, 2013.

[31] Kathrin Dentler, Annette ten Teije, Nicolette de Keizer, and Ronald Cornet. Barriers to the reuse of routinely recorded clinical data: a field report. *Studies in health technology and informatics*, 192:313–317, 2013.

[32] G F Cooper, B G Buchanan, M Kayaalp, M Saul, and J K Vries. Using computer modeling to help identify patient subgroups in clinical data repositories. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 180–184, 1998.

[33] J J Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, 37(4–5):394–403, 1998.

[34] James J Cimino. In defense of the desiderata. *Journal of biomedical informatics*, 39(3):299–306, 2005.

[35] Barry Smith. New desiderata for biomedical terminologies. In *Ontologies and Biomedical Informatics, Conference of the International Medical Informatics Association*, 2005.

[36] J. J. Cimino and X. Zhu. The practical impact of ontologies on biomedical informatics. *Methods of information in medicine*, 2006.

[37] Jonathan B L Bard and Seung Y Rhee. Ontologies in biology: design, applications and future challenges. *Nature reviews. Genetics*, 5(3):213–222, 2004.

[38] Thomas R. Gruber. *Formal Ontology in Conceptual Analysis and Knowledge Representation*, chapter Toward Principles for the Design of Ontologies used for Knowledge Sharing. Kluwer Academic Publishers, 1993.

[39] B Chandrasekaran, JR Josephson, and VR Benjamins. What are ontologies and why do we need them ? *IEEE Intelligent Systems*, 14(1):20–26, 1999.

[40] Anita Burgun. Desiderata for domain reference ontologies in biomedicine. *Journal of Biomedical Informatics*, 39:307–313, 2006.

[41] Stefan Schulz, Laszlo Balkanyi, Ronald Cornet, and Olivier Bodenreider. From concept representations to ontologies: A paradigm shift in health informatics? *Healthcare informatics research*, 19(4):235–242, 2013.

[42] André Q Andrade, Markus Kreuzthaler, Janna Hastings, Maria Krestyaninova, and Stefan Schulz. Requirements for semantic biobanks. *Studies in health technology and informatics*, 180:569–573, 2012.

[43] 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.

[44] Matthew E Holford, James P McCusker, Kei-Hoi Cheung, and Michael Krauthammer. A semantic web framework to integrate cancer omics data with biological knowledge. *BMC bioinformatics*, 13 Suppl 1:S10, 2012.

[45] Lars J Jensen and Peer Bork. Ontologies in quantitative biology: a basis for comparison, integration, and discovery. *PLoS biology*, 8(5):e1000374, 2010.

[46] Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Untangling statistical and biological models to understand network inference: the need for a genomics network ontology. *Frontiers in genetics*, 5:299, 2014.

[47] Yi Liu, Adrien Coulet, Paea LePendu, and Nigam H Shah. Using ontology-based annotation to profile disease research. *Journal of the American Medical Informatics Association : JAMIA*, 19(e1):e177–e186, 2012.

[48] Robert Hoehndorf, Michel Dumontier, and Georgios V Gkoutos. Evaluation of research in biomedical ontologies. *Briefings in bioinformatics*, 2012. In press.

[49] Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, and Mark A Musen. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(Web Server issue):W170–W173, 2009.

[50] Alan Ruttenberg, Tim Clark, William Bug, Matthias Samwald, Olivier Bodenreider, Helen Chen, Donald Doherty, Kerstin Forsberg, Yong Gao, Vipul Kashyap, June Kinoshita, Joanne Luciano, M. Scott Marshall, Chimezie Ogbuji, Jonathan Rees, Susie Stephens, Gwendolyn T. Wong, Elizabeth Wu, Davide Zaccagnini, Tonya Hongsermeier, Eric Neumann, Ivan Herman, and Kei-Hoi Cheung. Advancing translational research with the semantic web. *BMC Bioinformatics*, 8(3), 2007.

[51] Robert Stevens, Mikel Egaña Aranguren, Katy Wolstencroft, Ulrike Sattler, Nick Drummond, Matthew Horridge, and Alan Rector. Using OWL to model biological knowledge. *International Journal of Human Computer Studies*, 65(7):583–594, 2007.

[52] Mikel Egaña Aranguren, Sean Bechhofer, Phillip Lord, Ulrike Sattler, and Robert Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the gene ontology in OWL. *BMC bioinformatics*, 8:57, 2007.

[53] David P Hill, Nico Adams, Mike Bada, Colin Batchelor, Tanya Z Berardini, Heiko Dietze, Harold J Drabkin, Marcus Ennis, Rebecca E Foulger, Midori A Harris, Janna Hastings, Namrata S Kale, Paula de Matos, Christopher J Mungall, Gareth Owen, Paola Roncaglia, Christoph Steinbeck, Steve Turner, and Jane Lomax. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. *BMC genomics*, 14:513, 2013.

[54] Nicola Cannata, Michael Schröder, Roberto Marangoni, and Paolo Romano. A semantic web for bioinformatics: goals, tools, systems, applications. *BMC bioinformatics*, 9 Suppl 4:S1, 2008.

[55] Lennart J G Post, Marco Roos, M Scott Marshall, Roel van Driel, and Timo M Breit. A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data. *Bioinformatics (Oxford, England)*, 23(22):3080–3087, 2007.

[56] R Bellazzi. Big data and biomedical informatics: A challenging opportunity. *Yearbook of medical informatics*, 9(1), 2014. In press.

[57] Satya S. Sahoo, Olivier Bodenreider, Kelly Zeng, and Amit Sheth. An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information. In *Proceedings of the WWW2007 Workshop on Health Care and Life Sciences Data Integration for the Semantic Web*, 2007.

[58] María Taboada, Diego Martínez, Belén Pilo, Adriano Jiménez-Escrig, Peter N Robinson, and María J Sobrido. Querying phenotype-genotype relationships on patient datasets using semantic web technology: the example of cerebrotendinous xanthomatosis. *BMC medical informatics and decision making*, 12:78, 2012.

[59] Christian Bizer, Tom Heath, and Tim Berners Lee. Linked data–the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[60] Kevin M Livingston, Michael Bada, William A Baumgartner, and Lawrence E Hunter. Kabob: ontology-based semantic integration of biomedical databases. *BMC bioinformatics*, 16:126, 2015.

[61] 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.

[62] Kei-Hoi Cheung, H Robert Frost, M Scott Marshall, Eric Prud'hommeaux, Matthias Samwald, Jun Zhao, and Adrian Paschke. A journey to semantic web query federation in the life sciences. *BMC bioinformatics*, 10 Suppl 10:S10, 2009.

[63] Alison Callahan, José Cruz-Toledo, and Michel Dumontier. Ontology-based querying with Bio2RDF's linked open data. *Journal of biomedical semantics*, 4 Suppl 1:S1, 2013.

[64] Sarala M Wimalaratne, Jerven Bolleman, Nick Juty, Toshiaki Katayama, Michel Dumontier, Nicole Redaschi, Nicolas Le Novère, Henning Hermjakob, and Camille Laibe. SPARQL-enabled identifier conversion with identifiers.org. *Bioinformatics (Oxford, England)*, 31(11):1875–1877, 2015.

[65] Olivier Dameron, Bernard Gibaud, Anita Burgun, and Xavier Morandi. Towards a sharable numeric and symbolic knowledge base on cerebral cortex anatomy: lessons from a prototype. In *American Medical Informatics Association AMIA*, pages 185–189, 2002.

[66] Olivier Dameron, Bernard Gibaud, and Xavier Morandi. Numeric and symbolic representation of the cerebral cortex anatomy: Methods and preliminary results. *Surgical and Radiologic Anatomy*, 26(3):191–197, 2004.

[67] Olivier Dameron, Anita Burgun, Xavier Morandi, and Bernard Gibaud. Modelling dependencies between relations to insure consistency of a cerebral cortex anatomy knowledge base. In *Studies in Health technology and informatics*, pages 403–408, 2003.

[68] Olivier Dameron, Bernard Gibaud, and Mark Musen. Using semantic dependencies for consistency management of an ontology of brain-cortex anatomy. In *First International Workshop on Formal Biomedical Knowledge Representation KRMED04*, pages 30–38, 2004.

[69] Olivier Dameron, Mark A. Musen, and Bernard Gibaud. Using semantic dependencies for consistency management of an ontology of brain-cortex anatomy. *Artificial Intelligence in Medicine*, 39(3):217–225, 2007.

[70] Patrick Lambrix, Manal Habbouche, and Marta Pérez. Evaluation of ontology development tools for bioinformatics. *Bioinformatics (Oxford, England)*, 19(12):1564–1571, 2003.

[71] Robert Stevens, Chris Wroe, Sean Bechhofer, Phillip Lord, Alan Rector, and Carole Goble. Building ontologies in daml + oil. *Comparative and functional genomics*, 4(1):133–141, 2003.

[72] Holger Knublauch, Ray W. Fergerson, Natalya F. Noy, and Mark A. Musen. The protégé OWL plugin: An open development environment for semantic web applications. In *Proceeding of the Third International Semantic Web Conference (ISWC2004)*, volume 3298 of *Lecture Notes in Computer Science*, pages 229–243. Springer Berlin Heidelberg, 2004.

[73] Daniel L. Rubin, Olivier Dameron, and Mark A. Musen. Use of description logic classification to reason about consequences of penetrating injuries. In *American Medical Informatics Association Conference AMIA05*, pages 649–653, 2005.

[74] Daniel L. Rubin, Olivier Dameron, Yasser Bashir, David Grossman, Parvati Dev, and Mark A. Musen. Using ontologies linked with geometric models to reason about penetrating injuries. *Artificial Intelligence in Medicine*, 37(3):167–176, 2006.

[75] Olivier Dameron, Daniel L. Rubin, and Mark A. Musen. Challenges in converting frame-based ontology into OWL: the Foundational Model of Anatomy case-study. In *American Medical Informatics Association Conference AMIA05*, pages 181–185, 2005.

[76] Olivier Dameron and Julie Chabalier. Automatic generation of consistency constraints for an OWL representation of the FMA. In *10th International Protégé Conference*, 2007.

[77] Olivier Dameron. JOT: a scripting environment for creating and managing ontologies. In *7th International Protégé Conference*, 2004.

[78] Olivier Dameron, Élodie Roques, Daniel L. Rubin, Gwenaëlle Marquet, and Anita Burgun. Grading lung tumors using OWL-DL based reasoning. In *9th International Protégé Conference*, 2006.

[79] Gwenaëlle Marquet, Olivier Dameron, Stephan Saikali, Jean Mosser, and Anita Burgun. Grading glioma tumors using OWL-DL and NCI thesaurus. In *Proceedings of the American Medical Informatics Association Conference AMIA'07*, pages 508–512, 2007.

[80] Anand Kumar, Yum Lina Yip, Barry Smith, and Pierre Grenon. Bridging the gap between medical and bioinformatics: an ontological case study in colon carcinoma. *Computers in biology and medicine*, 36(7-8):694–711, 2005.

[81] Franck W. Hartel, Sherri de Corronado, Robert Dionne, Gilberto Fragoso, and Jennifer Golbeck. Modeling a description logic vocabulary for cancer research. *Journal of Biomedical Informatics*, 38:114–129, 2005.

[82] Julian Seidenberg and Alan Rector. Web ontology segmentation: analysis, classification and use. In *Proceedings of the World Wide Web Conference (WWW'06)*, pages 13–22, 2006.

[83] Julie Chabalier, Gwenaëlle Marquet, Olivier Dameron, and Anita Burgun. Enrichissement de la hiérarchie KEGG par l'exploitation de Gene Ontology. In *Workshop OGSB, JOBIM'06*, 2006.

[84] Julie Chabalier, Olivier Dameron, and Anita Burgun. Integrating disease and pathway ontologies. In *Proceedings of the ISMB conference, Poster Session*, 2007.

[85] Julie Chabalier, Olivier Dameron, and Anita Burgun. Integrating and querying disease and pathway ontologies: building an OWL model and using RDFS queries. In *Bio-Ontologies Special Interest Group, Intelligent Systems for Molecular Biology conference (ISMB'07)*, 2007.

[86] Julie Chabalier, Olivier Dameron, and Anita Burgun. Using knowledge about pathways as an organizing principle for disease ontologies. In *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM'07)*, 2007.

[87] Nicolas Lebreton, Olivier Dameron, Christophe Blanchet, and Julie Chabalier. Utilisation d'ontologies de tâches et de domaine pour la composition semi-automatique de services web bioinformatiques. In *Proceedings of the Journées Ouvertes de Biologie, Informatique et Mathématiques (Jobim 2008)*, 2008.

[88] Nicolas Lebreton, Christophe Blanchet, Julie Chabalier, and Olivier Dameron. Utilisation d'ontologies de tâches et de domaine pour la composition semi-automatique de services web bioinformatiques. In *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2009)*, 2009.

[89] Nicolas Lebreton, Christophe Blanchet, Daniela Barreiro Claro, Julie Chabalier, Anita Burgun, and Olivier Dameron. Verification of parameters semantic compatibility for semi-automatic web service composition: a generic case study. In *12th International Conference on Information Integration and Web-based Applications and Services (iiWAS2010)*, pages 845–848, 2010.

[90] Charles Bettembourg, Christian Diot, Anita Burgun, and Olivier Dameron. GO2PUB: Querying PubMed with semantic expansion of gene ontology terms. *Journal of biomedical semantics*, 3(1):7, 2012.

[91] Anita Burgun, Lynda Temal, Arnaud Rosier, Olivier Dameron, Philippe Mabo, Pierre Zweigenbaum, Regis Beuscart, David Delerue, and Henry Christine. Integrating clinical data with information transmitted by implantable cardiac defibrillators to support medical decision in telecardiology: the application ontology of the AKENATON project. In *Proceedings of the American Medical Informatics Association Conference AMIA*, page 992, 2010.

[92] Olivier Dameron, Pascal van Hille, Lynda Temal, Arnaud Rosier, Louise Deléger, Cyril Grouin, Pierre Zweigenbaum, and Anita Burgun. Comparison of OWL and SWRL-based ontology modeling strategies for the determination of pacemaker alerts severity. In *Proceedings of the American Medical Informatics Association Conference AMIA*, page 284, 2011.

[93] Pascal van Hille, Julie Jacques, Julien Taillard, Arnaud Rosier, David Delerue, Anita Burgun, and Olivier Dameron. Comparing Drools and ontology-based reasoning approaches for telecardiology decision support. *Studies in health technology and informatics*, 180:300–304, 2012.

[94] Olivier Dameron, Paolo Besana, Oussama Zekri, Annabel Bourdé, Anita Burgun, and Marc Cuggia. OWL model of clinical trial eligibility criteria compatible with partially-known information. In *Proceedings of the Semantic Web for Life Sciences workshop SWAT4LS2012*, 2012.

[95] Olivier Dameron, Paolo Besana, Oussama Zekri, Annabel Bourdé, Anita Burgun, and Marc Cuggia. OWL model of clinical trial eligibility criteria compatible with partially-known information. *Journal of Biomedical Semantics*, 4(1), 2013.

[96] Charles Bettembourg, Christian Diot, and Olivier Dameron. Semantic particularity measure for functional characterization of gene sets using Gene Ontology. *PLoS ONE*, 9(1):e86525, 2014.

[97] Charles Bettembourg, Christian Diot, and Olivier Dameron. Optimal threshold determination for interpreting semantic similarity and particularity: Application to the comparison of gene sets and metabolic pathways using GO and ChEBI. *PloS one*, 10(7):e0133579, 2015.

[98] Frederic Herault, Annie Vincent, Olivier Dameron, Pascale Le Roy, Pierre Cherel, and Marie Damon. The longissimus and semimembranosus muscles display marked differences in their gene expression profiles in pig. *PloS one*, 9(5):e96491, 2014.

[99] Sylvain Prigent, Guillaume Collet, Simon M Dittami, Ludovic Delage, Floriane Ethis de Corny, Olivier Dameron, Damien Eveillard, Sven Thiele, Jeanne Cambefort, Catherine Boyen, Anne Siegel, and Thierry Tonon. The genome-scale metabolic network of ectocarpus siliculosus (EctoGEM): a resource to study brown algal physiology and beyond. *The Plant journal : for cell and molecular biology*, 80(2):367–381, 2014.

[100] Jean Coquet, Geoffroy Andrieux, Jacques Nicolas, Olivier Dameron, and Nathalie Theret. Analysis of tgf-beta signalization pathway thanks to topological and semantic web methods. In *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2015), poster session*, 2015.

[101] Charles Bettembourg, Olivier Dameron, Anthony Bretaudeau, and Fabrice Legeai. Intégration et interrogation de réseaux de régulation génomique et post-génomique. In *Proceedings of the IN-OVIVE workshop (INtégration de sources/masses de données hétérogènes*

et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement), conférence IC (Ingénierie des Connaissances) PFIA, 2015.

[102] Philippe Finet, Régine Le Bouquin-Jeannès, and Olivier Dameron. La télémédecine dans la prise en charge des maladies chroniques [in french]. *Techniques Hospitalières*, (740), 2013.

[103] Philippe Finet, Régine Le Bouquin-Jeannès, Olivier Dameron, and Bernard Gibaud. Review of current telemedicine applications for chronic diseases: Toward a more integrated system? *IRBM*, 2015. In press.

[104] Philippe Finet, Bernard Gibaud, Olivier Dameron, and Régine Le Bouquin-Jeannès. Interopérabilité d'un système de capteurs en télémédecine. In *Proceedings of the Journées d'étude sur la Télésanté, UTC Compiègne*, 2015.

[105] Yann Rivault, Olivier Dameron, and Nolwenn Le Meur. Une infrastructure générique basée sur les apports du web sémantique pour l'analyse des bases médico-administratives. In *Proceedings of the IN-OVIVE workshop (INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement), conférence IC (Ingénierie des Connaissances) PFIA*, 2015.

[106] Nick Juty, Nicolas Le Novère, and Camille Laibe. Identifiers.org and miriam registry: community resources to provide persistent identification. *Nucleic acids research*, 40(Database issue):D580–D586, 2012.

[107] David P Hill, Barry Smith, Monica S McAndrews-Hill, and Judith A Blake. Gene ontology annotations: what they mean and where they come from. *BMC bioinformatics*, 9 Suppl 5:S2, 2008.

[108] Judith A Blake. Ten quick tips for using the gene ontology. *PLoS computational biology*, 9(11):e1003343, 2013.

[109] Kevin M Livingston, Michael Bada, Lawrence E Hunter, and Karin Verspoor. Representing annotation compositionality and provenance for the semantic web. *Journal of biomedical semantics*, 4:38, 2013.

[110] Gene Ontology Consortium. The gene ontology (GO)project in 2006. *Nucleic acids research*, 34(Database issue):D322–D326, 2006.

[111] Gwenaëlle Marquet, Jean Mosser, and Anita Burgun. Aligning biomedical ontologies using lexical methods and the UMLS: the case of disease ontologies. *Studies in health technology and informatics*, 124:781–786, 2006.

[112] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic acids research*, 32(Database issue):D277–D280, 2004.

[113] Xizeng Mao, Tao Cai, John G. Olyarchuk, and Liping Wei. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics*, 21(19):3787–3793, 2005.

[114] Michel Klein. Combining and relating ontologies : an analysis of problems and solutions. In *International Joint Conference on Artificial Intelligence IJCAI01*, 2001.

[115] Patrick Lambrix and He Tan. Sambo – a system for aligning and merging biomedical ontologies. *Journal of Web Semantics*, 4(3), 2006.

[116] I. Horrocks, P.F. Patel-Schneider, and F. van Harmelen. From $\mathcal{SHIQ}$ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26, 2003.

[117] Cornelius Rosse, Anand Kumar, Jose L V Mejino, Daniel L Cook, Landon T Detwiler, and Barry Smith. A strategy for improving and integrating biomedical ontologies. In *Proceedings, American Medical Informatics Association Fall Symposium AMIA2005*, pages 639–643, 2005.

[118] Olivier Dameron, Charles Bettembourg, and Nolwenn Le Meur. Measuring the evolution of ontology complexity: the Gene Ontology case study. *PLoS ONE*, 8(10):e75993, 2013.

[119] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–D270, 2004.

[120] Peter Haase, Jeen Broekstra, Andreas Eberhart, and Raphael Volz. A comparison of RDF query languages. In *Proceedings of the Third International Semantic Web Conference (ISWC2004)*, pages 502–517, 2004.

[121] Chris Wroe, Carole Goble, Antoon Goderis, Phillip Lord, and al. Recycling workflows and services through discovery and reuse: Research articles. In *Concurrency Computation : Practice and Experience*, volume 19, pages 181–194, 2007.

[122] Roman Vaculin and Katia Sycara. Monitoring execution of OWL-S web services. *Proceedings of OWL-S: Experiences and Directions Workshop, European Semantic Web Conference*, 2007.

[123] Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, and al. Taverna: a tool for building and running workflows of services. In *Nucleic Acids Research*, volume 34, pages W729–32, 2006.

[124] Sooyoung Yoo and Jinwook Choi. On the query reformulation technique for effective MEDLINE document retrieval. *Journal of biomedical informatics*, 43(5):686–693, 2010.

[125] Nicolas Griffon, Wiem Chebil, Laetitia Rollin, Gaetan Kerdelhue, Benoit Thirion, Jean-François Gehanno, and Stéfan Jacques Darmoni. Performance evaluation of unified medical language system's synonyms expansion to query PubMed. *BMC medical informatics and decision making*, 12:12, 2012.

[126] Zhiyong Lu, Won Kim, and W John Wilbur. Evaluation of query expansion using MeSH in PubMed. *Information retrieval*, 12(1):69–80, 2009.

[127] Zhiyong Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database : the journal of biological databases and curation*, 2011:baq036, 2011.

[128] Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui, Sylvain Gaudan, Mark Riethoven, and Peter Stoehr. EBIMed–text crunching to gather facts for proteins from medline. *Bioinformatics (Oxford, England)*, 23(2):e237–e244, 2007.

[129] Yasunori Yamamoto and Toshihisa Takagi. Biomedical knowledge navigation by literature clustering. *Journal of biomedical informatics*, 40(2):114–130, 2006.

[130] Yoshimasa Tsuruoka, Makoto Miwa, Kaisei Hamamoto, Jun'ichi Tsujii, and Sophia Ananiadou. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics (Oxford, England)*, 27(13):i111–i119, 2011.

[131] A Bajpai, S Davuluri, H Haridas, G Kasliwal, H Deepti, KS Sreelakshmi, DS Chandrashekar, P Bora, M Farouk, N Chitturi, V Samudyata, KP ArunNehru, and K Acharya. In search of the right literature search engine(s). *Nature precedings*, page http://dx.doi.org/10.1038/npre.2011.2101.3, 2011.

[132] Michael Muin, Paul Fontelo, Fang Liu, and Michael Ackerman. SLIM: an alternative web interface for MEDLINE/PubMed searches - a preliminary study. *BMC medical informatics and decision making*, 5:37, 2005.

[133] Bhanu C Vanteru, Jahangheer S Shaik, and Mohammed Yeasin. Semantically linking and browsing PubMed abstracts with gene ontology. *BMC genomics*, 9 Suppl 1:S10, 2008.

[134] J Bhogal, A Macfarlane, and P Smith. A review of ontology-based query expansion. *Information Processing and Management*, 43(4):866–886, 2007.

[135] Roberto Navigli and Paola Velardi. An analysis of ontology-based query expansion strategies. In *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia*, 2003.

[136] Sérgio Matos, Joel P Arrais, Joao Maia-Rodrigues, and José Luis Oliveira. Concept-based query expansion for retrieving gene related publications from MEDLINE. *BMC bioinformatics*, 11:212, 2010.

[137] M A Harris, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, G M Rubin, J A Blake, C Bult, M Dolan, H Drabkin, J T Eppig, D P Hill, L Ni, M Ringwald, R Balakrishnan, J M Cherry, K R Christie, M C Costanzo, S S Dwight, S Engel, D G Fisk, J E Hirschman, E L Hong, R S Nash, A Sethuraman, C L Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, S Y Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, E M Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N de la Cruz, P Tonellato, P Jaiswal, T Seigfried, R White, and Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic acids research*, 32(Database issue):D258–D261, 2004.

[138] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic acids research*, 32(Database issue):D262–D266, 2004.

[139] Andreas Doms and Michael Schroeder. Gopubmed: exploring pubmed with the gene ontology. *Nucleic acids research*, 33(Web Server issue):W783–W786, 2005.

[140] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–515, 2008.

[141] 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(Web Server issue):W541–W545, 2011.

[142] Manuel Salvadores, Matthew Horridge, Paul R Alexander, Ray W Fergerson, Mark A Musen, and Natalya F Noy. Using SPARQL to query Bioportal ontologies and metadata. In *Proceedings of the International Semantic Web Conference ISWC 2012*, volume 7650 of *Lecture Notes in Computer Science*, pages 180–195, 2012.

[143] Nigam H Shah, Tyler Cole, and Mark A Musen. Chapter 9: Analyses using disease ontologies. *PLoS computational biology*, 8(12):e1002827, 2012.

[144] Andreas Heßand Nicholas Kushmerick. Learning to attach metadata to web services. In D. et al. Fensel, editor, *Proceedings of the Second International Semantic Web Conference (ISWC2003)*, pages 258–273, 2003.

[145] Sherri de Coronado, Margaret W Haber, Nicholas Sioutos, Mark S Tuttle, and Lawrence W Wright. Nci thesaurus: using science-based terminology to integrate cancer research results. *Studies in health technology and informatics*, 107(Pt 1):33–37, 2004.

[146] D. Nardi, R. J. Brachman, F. Baader, W. Nutt, F. M. Donini, U. Sattler, D. Calvanese, R. Mölitor, G. De Giacomo, R. Küsters, F. Wolter, D. L. McGuinness, P. F. Patel-Schneider, R. Möller, V. Haarslev, I. Horrocks, A. Borgida, C. Welty, A. Rector, E. Franconi, M. Lenzerini, and R. Rosati. *The Description Logics Handbook : Theory, Implementation and Applications*. Cambridge University Press, 2003.

[147] Franz Baader, Ian Horrocks, and Ulrike Sattler. Description logics as ontology languages for the semantic web. In Dieter Hutter and Werner Stephan, editors, *Festschrift in honor of Jörg Siekmann*, Lecture Notes in Artificial Intelligence, 2003.

[148] Dieter Fensel, Ian Horrocks, Franck van Harmelen, Stefan Decker, Michael Erdmann, and Michel Klein. OIL in a nutshell. In *Knowledge Acquisition, Modeling and Management*, pages 1–16, 2000.

[149] Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. OWL 2: the next step for OWL. *Journal of Web Semantics*, 6(4):309–322, 2008.

[150] C. Rosse, J.L. Mejino, R. Modayur, B.R.and Jakobovits, K.P. Hinshaw, and J.F. Brinkley. Motivation and organizational principles for anatomical knowledge representation: The digital anatomist symbolic knowledge base. *Journal of the American Medical Informatics Association*, 5(1):17–40, Jan/Feb 1998.

[151] C. Rosse and J.L.V Mejino. A reference ontology for bioinformatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36:478–500, 2003.

[152] N.F. Noy, M.A. Musen, J.L.V. Mejino, and C. Rosse. Pushing the envelope: Challenges in a frame-based representation of human anatomy. *Data and Knowledge Engineering Journal*, 48(3):335–359, 2004.

[153] Jennifer Golbeck, Gilberto Fragoso, Franck Hartel, Jim Hendler, Jim Oberthaler, and Bijan Parsia. The national cancer institute's thésaurus and ontology. *Journal of Web Semantics*, 1(1):75–80, 2003.

[154] A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, and C. Wroe. Owl pizzas: Practical experience in teaching owl-dl: Common errors and common patterns. In *Proceedings of the European Conference on Knowledge Acquisition (EKAW-2004)*, pages 63–81, 2004.

[155] N. Noy and A. Rector. Defining n-ary relations on the semantic web: Use with individuals. In *W3C Techn. Report.*, 2004. http://www.w3.org/TR/swbp-n-aryRelations/.

[156] Christine Golbreich, Songmao Zhang, and Olivier Bodenreider. The foundational model of anatomy in OWL: Experience and perspectives. *Web Semantics*, 4(3):181–195, 2006.

[157] Songmao Zhang, Olivier Bodenreider, and Christine Golbreich. Experience in reasoning with the foundational model of anatomy in OWL DL. In *Proceedings of the Pacific Symposium on Biocomputing*, number 11, pages 200–211, 2006.

[158] Natalya F. Noy and Daniel L. Rubin. Translating the foundational model of anatomy into owl. Technical report, Stanford Medical Informatics, 2007.

[159] Daniel L. Cook, Jose L.V. Mejino, and Cornelius Rosse. Evolution of a foundational model of physiology: Symbolic representation for functional bioinformatics. In *Proceedings of the MEDINFO'04 Conference*, pages 336–340, 2004.

[160] IJ Haimowitz, RS Patil, and Szolovits P. Representing medical knowledge in a terminological language is difficult. In *Proceedings of Annual Symposium on Computational Applications in Medical Care*, pages 101–105, 1988.

[161] Christine Golbreich. Combining rule and ontology reasoners for the semantic web. In *Proceedings of Rules and Rule Markup Languages for the Semantic Web*, number 3323, pages 6–22, 2004.

[162] W Backman, D Bendel, and R Rakhit. The telecardiology revolution: improving the management of cardiac disease in primary care. *J R Soc Med.*, 103(11):442–6, 2010.

[163] K Nikus, J Lähteenmäki, P Lehto, and M Eskola. The role of continuous monitoring in a 24/7 telecardiology consultation service–a feasibility study. *J Electrocardiol.*, 42(6):473–80, 2009.

[164] CT Lin, KC Chang, CL Lin, CC Chiang, SW Lu, SS Chang, BS Lin, HY Liang, RJ Chen, YT Lee, and LW Ko. An intelligent telecardiology system using a wearable and wireless ecg to detect atrial fibrillation. *IEEE Trans Inf Technol Biomed.*, 14(3):726–33, 2010.

[165] P Rubel, J Fayn, G Nollo, D Assanelli, B Li, L Restier, S Adami, S Arod, H Atoui, M Ohlsson, L Simon-Chautemps, D Télisson, C Malossi, GL Ziliani, A Galassi, L Edenbrandt, and P Chevalier. Toward personal eHealth in cardiology. results from the EPI-MEDICS telemedicine project. *J Electrocardiol.*, 38(4 suppl):100–6, 2005.

[166] GY Lip, R Nieuwlaat, R Pisters, DA Lane, and HJ Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–72, 2010.

[167] European Heart Rhythm Association and European Association for Cardio-Thoracic Surgery, AJ Camm, P Kirchhof, GY Lip, U Schotten, I Savelieva, S Ernst, IC Van Gelder, N Al-Attar, G Hindricks, B Prendergast, H Heidbuchel, O Alfieri, A Angelini, D Atar, P Colonna, R De Caterina, J De Sutter, A Goette, B Gorenek, M Heldal, SH Hohloser, P Kolh, JY Le Heuzey, P Ponikowski, and FH Rutten. Guidelines for the management of atrial fibrillation: the task force for the management of atrial fibrillation of the european society of cardiology (ESC). *Eur Heart J.*, 31(19):2369–429, 2010.

[168] BC Grau, I Horrocks, B Motik, B Parsia, P Patel-Schneider, and U Sattler. OWL 2: The next step for OWL. *Journal of Web Semantics*, 6:309–22, 2008.

[169] I Horrocks, PF Patel-Schneider, H Boley, S Tabet, B Grosof, and M Dean. SWRL: A semantic web rule language combining OWL and RuleML. Technical report, W3C submission, 2004.

[170] GY Lip and JL Halperin. Improving stroke risk stratification in atrial fibrillation. *Am J Med.*, 3(6):484–8, 2010.

[171] E Sirin, B Parsia, BC Grau, A Kalyanpur, and Y Katz. Pellet: A practical OWL-DL reasoner. *Journal of Web Semantics*, 5:51–3, 2007.

[172] Huanying Helen Gu, Duo Wei, Jose L V Mejino, and Gai Elhanan. Relationship auditing of the fma ontology. *Journal of biomedical informatics*, 42(3):550–557, 2009.

[173] Simon Jupp, Robert Stevens, and Robert Hoehndorf. Logical gene ontology annotations (GOAL): exploring gene ontology annotations with owl. *Journal of biomedical semantics*, 3 Suppl 1:S3, 2012.

[174] Wiktoria Golik, Olivier Dameron, Jérôme Bugeon, Alice Fatet, Isabelle Hue, Catherine Hurtaud, Matthieu Reichstadt, Marie-Christine Salaün, Jean Vernet, Léa Joret, Frédéric Papazian, Claire Nédellec, and Pierre-Yves Le Bail. ATOL: the multi-species livestock trait ontology. In *Proceedings of the 6th Metadata and Semantics Research Conference MTSR*, 2012.

[175] Isabelle Hue, Jérôme Bugeon, Olivier Dameron, Alice Fatet, Catherine Hurtaud, Léa Joret, Marie-Christine Meunier-Salaün, Claire Nédellec, Matthieu Reichstadt, Jean Vernet, and Pierre-Yves Le Bail. ATOL and EOL ontologies, steps towards embryonic phenotypes shared worldwide? In *Proceedings of the 4th Mammalian Embryo Genomics Meeting, October 2013, Quebec City*, volume 149 of *Animal Reproduction Science*, page 99, 2014.

[176] Pierre-Yves Le Bail, Jérôme Bugeon, Olivier Dameron, Alice Fatet, Wiktoria Golik, Jean-François Hocquette, Catherine Hurtaud, Isabelle Hue, Catherine Jondreville, Léa Joret, Marie-Christine Meunier-Salaün, Jean Vernet, Claire Nedellec, Matthieu Reichstadt, and Philippe Chemineau. Un langage de référence pour le phénotypage des animaux d'élevage : l'ontologie ATOL. *Production Animale*, 27(3):195–208, 2014.

[177] Christine Golbreich, Matthew Horridge, Ian Horrocks, Boris Motik, and Rob Shearer. OBO and OWL: Leveraging semantic web technologies for the life sciences. In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007)*, volume 4825 of *Lecture Notes in Computer Science*, pages 169–182, 2007.

[178] FG David and BJ Mc Carthy. Epidemiology of brain tumors. *Curr Opin Neurol.*, 13:635–640, 2000.

[179] P Kleihues, DN Louis, BW Scheithauer, LB Rorke, G Reifenberger, PC Burger, and WK Cavenee. The WHO classification of tumors of the nervous system. *J Neuropathol Exp Neurol.*, 61(3):215–225, 2002.

[180] Marc Cuggia, Paolo Besana, and David Glasspool. Comparing semi-automatic systems for recruitment of patients to clinical trials. *International journal of medical informatics*, 80(6):371–388, 2011.

[181] Marc Cuggia, Jean-Charles Dufour, Paolo Besana, Olivier Dameron, Regis Duvauferrier, Dominique Fieschi, Catherine Bohec, Annabel Bourdé, Laurent Charlois, Cyril Garde, Isabelle Gibaud, Jean-Francois Laurent, Oussama Zekri, and Marius Fieschi. ASTEC: A system for automatic selection of clinical trials. In *Proceedings of the American Medical Informatics Association Conference AMIA*, page 1729, 2011.

[182] Paolo Besana, Marc Cuggia, Oussama Zekri, Annabel Bourdé, and Anita Burgun. Using semantic web technologies for clinical trial recruitment. In *9th International Semantic Web Conference (ISWC2010)*, 2010.

[183] Derek Corrigan, Adel Taweel, Tom Fahey, Theodoras Arvanitis, and Brendan Delaney. An ontolological treatment of clinical prediction rules implementing the alvarado score. *Studies in health technology and informatics*, 186:103–107, 2013.

[184] Jean-Francois Ethier, Olivier Dameron, Vasa Curcin, Mark M. McGilchrist, Robert A. Verheij, Theodoros N. Arvanitis, Adel Taweel, Brendan C. Delaney, and Anita Burgun. A unified structural/terminological framework based on LexEVS: application to TRANSFoRm. *Journal of the American Medical Informatics Association*, 20(5):986–994, 2013.

[185] L Ohno-Machado, E Parra, S B Henry, S W Tu, and M A Musen. AIDS2: a decision-support tool for decreasing physicians' uncertainty regarding patient eligibility for HIV treatment protocols. *Proceedings Symposium on Computer Applications in Medical Care*, pages 429–433, 1993.

[186] Xiujie Chen, Ruizhi Yang, Jiankai Xu, Hongzhe Ma, Sheng Chen, Xiusen Bian, and Lei Liu. A sensitive method for computing go-based functional similarities among genes with 'shallow annotation'. *Gene*, 509(1):131–135, 2012.

[187] Sidahmed Benabderrahmane, Malika Smail-Tabbone, Olivier Poch, Amedeo Napoli, and Marie-Dominique Devignes. Intelligo: a new vector-based semantic similarity measure including annotation origin. *BMC bioinformatics*, 11(1):588, 2010.

[188] Marcus C Chibucos, Christopher J Mungall, Rama Balakrishnan, Karen R Christie, Rachael P Huntley, Owen White, Judith A Blake, Suzanna E Lewis, and Michelle Giglio. Standardized description of scientific evidence using the evidence ontology (eco). *Database : the journal of biological databases and curation*, 2014, 2014. In press.

[189] Frederic B Bastian, Marcus C Chibucos, Pascale Gaudet, Michelle Giglio, Gemma L Holliday, Hong Huang, Suzanna E Lewis, Anne Niknejad, Sandra Orchard, Sylvain Poux, Nives Skunca, and Marc Robinson-Rechavi. The confidence information ontology: a step towards a standard for asserting confidence in annotations. *Database : the journal of biological databases and curation*, 2015, 2015. In press.

[190] Wei-Nchih Lee, Nigam Shah, Karanjot Sundlass, and Mark Musen. Comparison of ontology-based semantic-similarity measures. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 384–388, 2008.

[191] Michael F Ochs, Aidan J Peterson, Andrew Kossenkov, and Ghislain Bidaut. Incorporation of gene ontology annotations to enhance microarray data analysis. *Methods in molecular biology (Clifton, N.J.)*, 377:243–254, 2007.

[192] Kristian Ovaska, Marko Laakso, and Sampsa Hautaniemi. Fast gene ontology based clustering for microarray experiments. *BioData mining*, 1(1):11, 2008.

[193] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics (Oxford, England)*, 23(10):1274–1281, 2007.

[194] Rafal Kustra and Adam Zagdanski. Incorporating gene ontology in clustering gene expression data. In *CBMS*, pages 555–563. IEEE Computer Society, 2006.

[195] Nadia Bolshakova, Francisco Azuaje, and Pádraig Cunningham. A knowledge-driven approach to cluster validity assessment. *Bioinformatics (Oxford, England)*, 21(10):2546–2547, 2005.

[196] Billy Chang, Rafal Kustra, and Weidong Tian. Functional-network-based gene set analysis using gene-ontology. *PloS one*, 8(2):e55635, 2013.

[197] Catia Pesquita, Daniel Faria, Hugo Bastos, Antònio EN Ferreira, Falcaon André O, and Francisco M Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, 2008.

[198] Catia Pesquita, Daniel Faria, André O Falcão, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443, 2009.

[199] Emmanuel Blanchard, Mounira Harzallah, and Pascale Kuntz. A generic framework for comparing semantic similarities on a subsumption hierarchy. In *ECAI2008 - 18th European Conference on Artificial Intelligence*, pages 20–24, 2008.

[200] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[201] Dekang Lin. An information-theoric definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, 1998.

[202] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, page 9008, 1997.

[203] G. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(1):39–41, 1995.

[204] Catia Pesquita, Daniel Faria, Hugo Bastos, André O Falcão, and Francisco M Couto. Evaluating go-based semantic similarity measures. In *Proc. 10th Annual Bio-Ontologies Meeting*, pages 37–40, 2007.

[205] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[206] Brendan Sheehan, Aaron Quigley, Benoit Gaudin, and Simon Dobson. A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, 9(1):468, 2008.

[207] Francisco M Couto, Mário J Silva, and Pedro M Coutinho. Measuring semantic similarity between gene ontology terms. *Data and Knowledge Engineering*, 61(1):137–152, 2007.

[208] Bo Jin and Xinghua Lu. Identifying informative subsets of the gene ontology with information bottleneck methods. *Bioinformatics (Oxford, England)*, 26(19):2445–2451, 2010.

[209] Jesse Gillis and Paul Pavlidis. Assessing identity, redundancy and confounds in gene ontology annotations over time. *Bioinformatics (Oxford, England)*, 2013. In press.

[210] Gang Chen, Jianhuang Li, and Jianxin Wang. Evaluation of gene ontology semantic similarities on protein interaction datasets. *International journal of bioinformatics research and applications*, 9(2):173–183, 2013.

[211] R Rada, H Mili, E Bicknell, and M Blettner. Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[212] Viktor Pekar and Steffen Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of COLING*, 2002.

[213] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proc. of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, 1994.

[214] Jill Cheng, Melissa Cline, John Martin, David Finkelstein, Tarif Awad, David Kulp, and Michael A Siani-Rose. A knowledge-based clustering algorithm driven by Gene Ontology. *Journal of biopharmaceutical statistics*, 14(3):687–700, 2004.

[215] Marco A Alvarez and Changhui Yan. A graph-based semantic similarity measure for the gene ontology. *Journal of bioinformatics and computational biology*, 9(6):681–695, 2011.

[216] Norberto Díaz-Díaz and Jesús S Aguilar-Ruiz. Go-based functional dissimilarity of gene sets. *BMC bioinformatics*, 12:360, 2011.

[217] Gaston K Mazandu and Nicola J Mulder. Dago-fun: tool for gene ontology-based functional analysis using term information content measures. *BMC bioinformatics*, 14(1):284, 2013.

[218] Steffen Grossmann, Sebastian Bauer, Peter N Robinson, and Martin Vingron. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. *Bioinformatics (Oxford, England)*, 23(22):3024–3031, 2007.

[219] Sebastian Klie, Marek Mutwil, Staffan Persson, and Zoran Nikoloski. Inferring gene functions through dissection of relevance networks: interleaving the intra- and inter-species views. *Molecular bioSystems*, 8(9):2233–2241, 2012.

[220] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2008.

[221] Roland Barriot, David J Sherman, and Isabelle Dutour. How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. *BMC bioinformatics*, 8:332, 2007.

[222] Miranda D Stobbe, Gerbert A Jansen, Perry D Moerland, and Antoine H C van Kampen. Knowledge representation in metabolic pathway databases. *Briefings in bioinformatics*, 2012. In press.

[223] Troy Hawkins, Meghana Chitale, and Daisuke Kihara. Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by pfp. *BMC bioinformatics*, 11:265, 2010.

[224] Zhixia Teng, Maozu Guo, Xiaoyan Liu, Qiguo Dai, Chunyu Wang, and Ping Xuan. Measuring gene functional similarity based on group-wise comparison of go terms. *Bioinformatics (Oxford, England)*, 2013. In press.

[225] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[226] Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, 21(16):3448–3449, 2005.

[227] Qi Zheng and Xiu-Jie Wang. GOEAST: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic acids research*, 36(Web Server issue):W358–W363, 2008.

[228] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics (Oxford, England)*, 25(8):1091–1093, 2009.

[229] Da Wei Huang, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, 8(9):R183, 2007.

[230] Erich J Baker, Jeremy J Jay, Jason A Bubier, Michael A Langston, and Elissa J Chesler. GeneWeaver: a web-based system for integrative functional genomics. *Nucleic acids research*, 40(Database issue):D1067–D1076, 2011.

[231] Bing Zhang, Denise Schmoyer, Stefan Kirov, and Jay Snoddy. Gotree machine (gotm): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC bioinformatics*, 5:16, 2004.

[232] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2011.

[233] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1):25–29, 2000.

[234] Jing Wang, Xianxiao Zhou, Jing Zhu, Yunyan Gu, Wenyuan Zhao, Jinfeng Zou, and Zheng Guo. Go-function: deriving biologically relevant functions from statistically significant functions. *Briefings in bioinformatics*, 13(2):216–227, 2011.

[235] Wyatt T Clark and Predrag Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics (Oxford, England)*, 29(13):i53–i61, 2013.

[236] Satoshi Shibata, Mitsuho Sasaki, Takashi Miki, Akira Shimamoto, Yasuhiro Furuichi, Jun Katahira, and Yoshihiro Yoneda. Exportin-5 orthologues are functionally divergent among species. *Nucleic acids research*, 34(17):4711–4721, 2006.

[237] David Croft, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(Database issue):D691–D697, 2010.

[238] 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(Database issue):D344–D350, 2007.

[239] Graeme D Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann-whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

[240] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.

[241] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic acids research*, 41(Database issue):D8–D20, 2012.

[242] Thomas Kelder, Martijn P van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. Wikipathways: building research communities on biological pathways. *Nucleic acids research*, 2011. In press.

[243] Ron Caspi, Hartmut Foerster, Carol A. Fulcher, Rebecca Hopkinson, John Ingraham, Pallavi Kaipa, Markus Krummenacker, Suzanne Paley, John Pick, Seung Y. Rhee, Christophe Tissier, Peifen Zhang, and Peter D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, 35:D511–D516, 2006.

[244] Lean Strömbäck and Patrick Lambrix. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21(24):4401–4407, 2005.

[245] Alan Ruttenberg, Jonathan Rees, and Jeremy Zucker. What biopax communicates and how to extend owl to help it. In Bernardo Cuenca Grau, Pascal Hitzler, Conor Shankey, and Evan Wallace, editors, *Proceedings of the OWLED-06 Workshop on OWL: Experiences and Directions*, volume 216. CEUR-WS.org, 2006.

[246] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl andoanne Schaefer, Frank Schacherer, Irma Martinez-Flores, Zhenjun Hu, Veronica Jimenez-Jacinto, Geeta Joshi-Tope, Kumaran Kandasamy, Alejandra C Lopez-Fuentes, Elgar Mi, Huchler, Igor Rodchenkov, Andrea Splendiani, Sasha Tkachev, Jeremy Zucker, Gopal Gopinath, Harsha Rajasimha, Ranjani Ramakrishnan, Imran Shah, Nadia Syeand Anwar, Ozgün Babur, Michael Blinov, Erik Brauner, Dan Corwin, Sylva Donaldson, Frank Gibbons, Robert Goldberg, Peter Hornbeck, Peter Luna, Augustiy-Rust, Eric Neumann, Oliver Ruebenacker, Oliver Reubenacker, Matthias Samwald, Martijn van Iersel, Sarala Wimalaratne, Keith Allen, Michelle Braun, Burk andillo, Kei-Hoi Cheung, Kam Dahlquist, Andrew Finney, Marc

Gillespie, Elizabeth Glass, Li Gong, Robin Haw, Michael Honig, Olivier Hubaut, Shiva Kand Krupa, Martina Kutmon, Julie Leonard, Debbie Marks, David Merberg, Victoria Petri, Alex Pico, Dean Ravenscroft, Liya Ren, Nigam Shah, Sugot, Rebecca Tang, Ryan Whaley, Stan Letovksy, Kenneth H Buetow, Andrey Rzhetsky, Vincent Schachter, Bruno S Sobral, Ugur Dogrusoz, Shannon a Mirit McWeeney, Ewan Birney, Julio Collado-Vides, Susumu Goto, Michael Hucka, Nicolas Le Novère, Natalia Maltsev, Akhilesh Pandey, Paul Thomas, Edga Peter D Wingender, Chris Sander, and Gary D Bader. The BioPAX community standard for pathway data sharing. *Nature biotechnology*, 28(9):935–942, 2010.

[247] Donny Soh, Difeng Dong, Yike Guo, and Limsoon Wong. Consistency, comprehensiveness, and compatibility of pathway databases. *BMC bioinformatics*, 11:449, 2010.

[248] Tomer Altman, Michael Travers, Anamika Kothari, Ron Caspi, and Peter D Karp. A systematic comparison of the metacyc and kegg pathway databases. *BMC bioinformatics*, 14:112, 2013.

[249] Liam G Fearnley, Melissa J Davis, Mark A Ragan, and Lars K Nielsen. Extracting reaction networks from databases-opening pandora's box. *Briefings in bioinformatics*, 15(6):973–983, 2014.

[250] Nicole Redaschi and Consortium UniProt. UniProt in RDF: Tackling data integration and distributed annotation with the semantic web. In *3rd International Biocuration Conference*, 2009. Available from Nature Precedings http://dx.doi.org/10.1038/npre.2009.3193.1.

[251] Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, Sarala M Wimalaratne, Maria Martin, Nicolas Le Novère, Helen Parkinson, Ewan Birney, and Andrew M Jenkinson. The ebi rdf platform: linked open data for the life sciences. *Bioinformatics (Oxford, England)*, 30(9):1338–1339, 2014.

[252] Gang Fu, Colin Batchelor, Michel Dumontier, Janna Hastings, Egon Willighagen, and Evan Bolton. PubChemRDF: towards the semantic annotation of pubchem compound and substance databases. *Journal of cheminformatics*, 7:34, 2015.

[253] Thomas Kelder, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. Finding the right questions: exploratory pathway analysis to enhance biological discovery in large datasets. *PLoS biology*, 8(8):e1000472, 2010.

[254] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8 Suppl 2:I1, 2014.

[255] Gautier Defossez, Alexandre Rollet, Olivier Dameron, and Pierre Ingrand. Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer. *BMC Medical Informatics and Decision Making*, 14(1):24, 2014.

[256] Joanne S. Luciano and Robert D. Stevens. e-Science and biological pathway semantics. *BMC Bioinformatics*, 8(Suppl 3):S3, 2007.

[257] Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34(Web Server issue):W729–W732, 2006.

[258] Carole A Goble, Jiten Bhagat, Sergejs Aleksejevs, Don Cruickshank, Danius Michaelides, David Newman, Mark Borkum, Sean Bechhofer, Marco Roos, Peter Li, and David De Roure. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(Web Server issue):W677–W682, 2010.

[259] Jeremy Goecks, Anton Nekrutenko, James Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, 2010.

[260] Daniel Blankenberg, James E Johnson, Galaxy Team, James Taylor, and Anton Nekrutenko. Wrangling galaxy's reference data. *Bioinformatics (Oxford, England)*, 30(13):1917–1919, 2014.

[261] Alejandra González-Beltrán, Eamonn Maguire, Susanna-Assunta Sansone, and Philippe Rocca-Serra. linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC bioinformatics*, 15 Suppl 14:S4, 2014.

[262] Jon Ison, Matús Kalas, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, and Peter Rice. Edam: An ontology of bioinformatics operations, types of data and identifiers, topics, and formats. *Bioinformatics (Oxford, England)*, 2013. In press.

[263] François Moreews, Yvan Le Bras, Olivier Dameron, Cyril Monjeaud, and Olivier Collin. Integrating galaxy workflows in a metadata management environment. In *Galaxy Community Conference GCC2014, Proceedings*, 2014.

[264] Sébastien Ferré. Expressive and scalable query-based faceted search over SPARQL endpoints. In *International Semantic Web Conference (ISWC)*, pages 438–453, 2014.

[265] Samuel M D Seaver, Christopher S Henry, and Andrew D Hanson. Frontiers in metabolic reconstruction and modeling of plant genomes. *Journal of experimental botany*, 63(6):2247–2258, 2012.

[266] Cristiana Gomes de Oliveira Dal'Molin and Lars Keld Nielsen. Plant genome-scale metabolic reconstruction and modelling. *Current opinion in biotechnology*, 24(2):271–277, 2012.

[267] Shira Mintz-Oron, Sagit Meir, Sergey Malitsky, Eytan Ruppin, Asaph Aharoni, and Tomer Shlomi. Reconstruction of arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 109(1):339–344, 2011.

[268] Seongwon Seo and Harris A Lewin. Reconstruction of metabolic pathways for the cattle genome. *BMC systems biology*, 3:33, 2009.

[269] Baikang Pei and Dong-Guk Shin. Reconstruction of biological networks by incorporating prior knowledge into bayesian network models. *Journal of computational biology*, 19(12):1324–1334, 2012.

[270] Seyedsasan Hashemikhabir, Eyup Serdar Ayaz, Yusuf Kavurucu, Tolga Can, and Tamer Kahveci. Large-scale signaling network reconstruction. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 9(6):1696–1708, 2012.

[271] Chen Li, Maria Liakata, and Dietrich Rebholz-Schuhmann. Biological network extraction from scientific literature: state of the art and challenges. *Briefings in bioinformatics*, 15(5):856–877, 2013.

[272] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012.

[273] Ron Caspi, Kate Dreher, and Peter D Karp. The challenge of constructing, classifying and representing metabolic pathways. *FEMS microbiology letters*, 345(2):85–93, 2013.

[274] John P Rooney, Ashish Patil, Fraulin Joseph, Lauren Endres, Ulrike Begley, Maria R Zappala, Richard P Cunningham, and Thomas J Begley. Cross-species functionome analysis identifies proteins associated with dna repair, translation and aerobic respiration as conserved modulators of uv-toxicity. *Genomics*, 97(3):133–147, 2010.

[275] Gamze Abaka, Türker Biyikoglu, and Cesim Erten. CAMPways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways. *Bioinformatics (Oxford, England)*, 29(13):i145–i153, 2013.

[276] Ron Caspi, Tomer Altman, Kate Dreher, Carol A Fulcher, Pallavi Subhraveti, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Quang Ong, Suzanne Paley, Anuradha Pujar, Alexander G Shearer, Michael Travers, Deepika Weerasinghe, Peifen Zhang, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 40(Database issue):D742–D753, 2011.

[277] Joan Massagué. Tgfbeta signalling in context. *Nature reviews. Molecular cell biology*, 13(10):616–630, 2012.

[278] Geoffroy Andrieux, Michel Le Borgne, and Nathalie Théret. An integrative modeling framework reveals plasticity of tgf-beta signaling. *BMC systems biology*, 8(1):30, 2014.

[279] Kemal Eren, Mehmet Deveci, Onur Küçüktunç, and Ümit V Çatalyürek. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in bioinformatics*, 14(3):279–292, 2013.

[280] Alessia Visconti, Francesca Cordero, and Ruggero G Pensa. Leveraging additional knowledge to support coherent bicluster discovery in gene expression data. *Intelligent Data Analysis*, 18(5):837–855, 2014.

[281] Rohit Gupta, Navneet Rao, and Vipin Kumar. Discovery of error-tolerant biclusters from noisy gene expression data. *BMC bioinformatics*, 12 Suppl 12:S1, 2011.

[282] Rui Henriques and Sara C Madeira. Bicpam: Pattern-based biclustering for biomedical data analysis. *Algorithms for molecular biology : AMB*, 9(1):27, 2014.

[283] Rudolf Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In I Rival, editor, *Ordered sets*, volume 83 of *NATO Advanced Study Institutes Series*, pages 445–470. Springer Netherlands, 1982.

[284] Sylvain Blachon, Ruggero G Pensa, Jérémy Besson, Céline Robardet, Jean-François Boulicaut, and Olivier Gandrillon. Clustering formal concepts to discover biologically relevant knowledge from gene expression data. *In silico biology*, 7(4-5):467–483, 2007.

[285] Santiago Videla, Carito Guziolowski, Federica Eduati, Sven Thiele, martin Gebser, Jacques Nicolas, Julio Saez-Rodriguez, Torsten Schaub, and Anne Siegel. Learning boolean logic models of signaling networks with ASP. *Theoretical Computer Science*, 2014. In press.

[286] Mehdi Kaytoue, Sergei O. Kuznetsov, Juraj Macko, and Amedeo Napoli. Biclustering meets triadic concept analysis. *Annals of Mathematics and Artificial Intelligence, Special Issue Post-proceedings of CLA 2011*, 70(1–2):55–79, 2014.

[287] Mehdi Kaytoue, Victor Codocedo, Jaume Baixeries, and Amedeo Napoli. Three interrelated fca methods for mining biclusters of similar values on columns. In Karell Bertet and Sebastian Rudolph, editors, *The Eleventh International Conference on Concept Lattices and their Applications (CLA 2014), Kosice Slovakia*, pages 243–254. CEUR Workshop Proceedings 1252, 2014.

[288] Aleksey Buzmakov, Sergei O. Kuznetsov, and Amedeo Napoli. Scalable estimates of stability. In Cynthia Vera Glodeanu, Mehdi Kaytoue, and Christian Sacarea, editors, *12th International Conference on Formal Concept Analysis (ICFCA 2014)*, Lecture Notes in Artificial Intelligence LNAI 8478, pages 157–172. Springer, 2014.

[289] Aleksey Buzmakov, Sergei O. Kuznetsov, and Amedeo Napoli. On evaluating interestingness measures for closed itemsets. In *7th European Starting AI Researcher Symposium (STAIRS-2014), Prague*, pages 71–80. IOS Press, 2014.

[290] Simon Andrews and Constantinos Orphanides. Analysis of large data sets using formal concept lattices. In Marzena Kryszkiewicz and Sergei Obiedkov, editors, *Proceedings of the 7th International Conference on Concept Lattices and Their Applications*, volume 672 of *CEUR Workshop Proceedings*, 2010.

[291] Ruggero G Pensa and Jean-François Boulicaut. Towards fault-tolerant formal concept analysis. In *Proceedings of AI*IA 2005: Advances in Artificial Intelligence*, volume 3673 of *Lecture Notes in Computer Science*, pages 212–223, 2005.

Olivier Dameron

Curriculum Vitæ

Identification

Nom patronymique : Dameron
Nom usuel : Dameron
Prénom : Olivier
Date de naissance : 23 octobre 1974
Grade : Maître de conférences – classe normale
Établissement : Université de Rennes 1
Section CNU : 65
ORCID 0000-0001-8959-7189

Domaine de recherche

Je développe des **méthodes basées sur les ontologies pour analyser des données** biomédicales. Cela fait intervenir des compétences en représentation des connaissances et en bioinformatique.

Mon approche consiste à exploiter des connaissances symboliques du domaine d'étude afin d'améliorer l'analyse de données qui sont en grandes quantités, complexes, fortement interdépendantes et incomplètes. Pour cela, j'utilise les technologies du **Web Sémantique** pour intégrer ces données qui sont souvent distribuées, et pour combiner différents types de raisonnement : déduction, classification, comparaison...

L'application principales concerne la caractérisation fonctionnelle et la comparaison de voies métaboliques et de voies de signalisation.

Déroulement de carrière

depuis septembre 2005 :	Maître de conférences, Université de Rennes 1.
janvier 2004 – juin 2005 :	Postdoctorant. Stanford Medical Informatics group, Université de Stanford (Californie, États-Unis d'Amérique). Responsable : Mark Musen.
octobre 2000 – décembre 2003 :	Doctorat Modélisation, représentation et partage de connaissances anatomiques sur le cortex cérébral – Université de Rennes 1, directeur : Bernard Gibaud.
1999 – 2000	DEA informatique médicale, Université de Rennes 1. 1er/14.
1998 – 1999	Service militaire.
1995 – 1998	Élève ingénieur INSA Rennes, département informatique.

Activités de responsabilités (administratives et en recherche sur toute la carrière)

Recherche

2015–présent :	Membre nommé de la section 65 du CNU
2015 :	Responsable du PEPS CNRS Fondements et applications de la science des données « confocal » (concepts formels, connaissances ontologies et étude de liaisons).
2014–présent :	Coordinateur du thème « Biologie-santé » de l'IRISA.

2007 : Jury de thèse de Sandrine Pawlicki (Approche bioinformatique des mécanismes d'agrégation et de polymérisation des protéines amyloïdes).

2005–2013 : Création et co-responsabilité du thème « Réseaux d'Expression Génétique : *in vivo*, *in vitro* et *in silico* » au sein de l'IFR 140.

2005 : Comité d'organisation 8^{eme} conf. internationale Protégé.

2004 : Comité d'organisation 7^{eme} conf. internationale Protégé.

Enseignement

Responsabilités de formations

depuis septembre 2012 : Co-responsable du master 2 « Bioinformatique et génomique ».

2008–2012 : Co-responsable du parcours « Bioinformatique » du master 2 « Modélisation des systèmes biologiques ».

2007–2012 : Responsable du master 1 recherche « Méthodes et Traitements de l'Information Biomédicale et Hospitalière ».

2010–2012 : Co-responsable du master 2 « Méthodes et Traitements de l'Information Biomédicale et Hospitalière ».

Responsabilités d'UE

depuis septembre 2005 : Responsable des UE « Bases de mathématiques et probabilité » et « Méthodes en informatique » du master 1 « Santé publique ».

depuis septembre 2006 : Responsable de l'UE « Méthodes Web avancé en biomédical » du master 2 recherche « Santé publique ».

depuis septembre 2006 : Responsable de l'UE « Principes de programmation et algorithmique » du master 1 « Bioinformatique et génomique ».

depuis septembre 2008 : Responsable de l'UE « Standardisation des connaissances et bio-ontologies » du master 2 « santé publique ».

depuis septembre 2011 : Responsable du module « eSanté » en troisième année de cycle ingénieur ESIR.

depuis septembre 2012 : Responsable de l'UE « Gestion de projet informatique » du master 1 « Bioinformatique et génomique ».

depuis septembre 2015 : Co-responsable de l'UE « Bioinformatique » ENS Rennes.

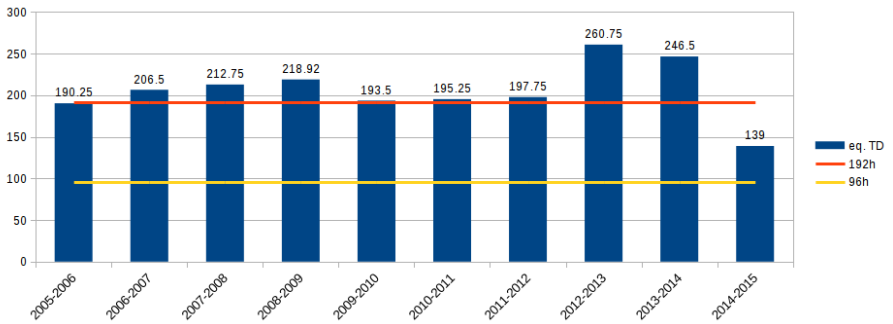

FIG. 1 – *Volume des enseignements annuels. La charge normale est de 192h équivalent TD. J'étais en demi-délégation à l'INRIA en 2014–2015 et 2015–2016 et ne devait donc que 96h equivalent TD lors de cette période.*

Encadrement

– Doctorats : 5 (dont 3 en cours)

- Stages ingénieurs : 1
- Stages master2 : 7
- Stages master1 : 9

mars – juin 2006 : Stage master 1 Bioinformatique : Élodie Roques.

janvier – juin 2006 : Stage master 2 recherche Informatique médicale : Ihssène Belhadj.

février – juin 2006 : Stage master 2 professionnel Traitement de l'information médicale et hospitalière : Nicolas Cottais

janvier – juin 2007 : Stage master 2 recherche Bioinformatique : Élodie Roques.

2007–2010 : Doctorat Nicolas Lebreton (bourse MENRT) : « Réalisation d'ontologies de tâches et de domaine en bioinformatique et utilisation de la sémantique pour l'appariement semi-automatique de Services Web ». Co-encadrement avec Anita Burgun.

avril – juin 2008 : Stage master 1 bioinformatique : Léa Joret.

janvier – juin 2009 : Stage master 2 recherche bioinformatique : Léa Joret.

avril – juin 2009 : Stage master 1 bioinformatique : Charles Bettembourg.

janvier – juin 2010 : Stage master 2 recherche bioinformatique : Charles Bettembourg.

juillet 2010 – mai 2011 : Stage ingénieur CNAM : Pascal van Hille.

2010 – 2013 : Doctorat Charles Bettembourg (bourse MENRT) : « Comparaison inter-espèces de voies métaboliques : application à l'étude du métabolisme des lipides chez le poulet, la souris et l'homme ». Co-encadrement avec Christian Diot (INRA).

avril – juin 2011 : Stage master 1 bioinformatique : Walid Bedhiafi.

avril – septembre 2011 : Stage master 2 CCI : Nicolas Schnell.

avril – juin 2012 : Stage master 1 Bioinformatique et génomique : Jérémy Rio.

avril – juin 2013 : Stage master 1 Bioinformatique et génomique : Ayité Kougbeadjo.

octobre 2013 – Doctorat Philippe Finet (ingénieur en CDI à la DSI du CHU Alençon-Mamers) : « Production et transmission des données de suivi des patients dans un contexte de télémédecine et intégration dans un système d'information pour l'aide à la décision ». Co-encadrement avec Régine Le Bouquin-Jeannes du LTSI.

avril – juin 2014 : Stage master 1 Bioinformatique et génomique : Dominique Mias-Lucquin.

avril – juin 2014 : Stage master 1 Bioinformatique et génomique : Loïc Bourgeois.

octobre 2014 – Doctorat Jean Coquet (bourse MENRT) : « Semantic-based reasoning for biological pathways analysis ». Co-encadrement avec Jacques Nicolas

mars – août 2015 : Stage master 2 Statistiques pour l'entreprise : Yann Rivault (co-encadré avec Nolwenn Le Meur, EHESP).

avril – juin 2015 : Stage master 1 Bioinformatique et génomique : Pierre Vignet.

octobre 2015 – Doctorat Yann Rivault (contrat ANSM) : « ». Co-encadrement avec Nolwenn Le Meur (EHESP)

Distinctions, rayonnement scientifique et relation avec le monde industriel

Distinctions

janvier – décembre 2004 : Lauréat bourse INRIA de stage postdoctoral à l'étranger.

2011 : L'article « Comparison of OWL and SWRL-based ontology modeling strategies for the determination of pacemaker alerts severity » a été sélectionné pour le *Best paper award* de la conférence AMIA (*American Medical Informatics Association*). Une version étendue a été soumise à un journal.

2014 – 2015 : demi délégation INRIA.

2015 – 2016 : demi délégation INRIA.

Rayonnement scientifique

2005–2011 :	Participation à l'animation des « Protégé Short Course » et « Protégé-OWL Short Course ». Il s'agit de sessions payantes de formation à destination d'un public d'industriels et d'académiques, organisées par le Stanford Medical Informatics group. J'ai participé à leur création leur de mon stage postdoctoral et j'ai ensuite été régulièrement invité pour animer une partie de ces formations jusqu'en 2011.
avril 2009 :	Séminaire invité LRI, Orsay.
mai 2010 :	Présentation BreizhJUG : *Introduction to the semantic Web*
octobre 2011 :	Journée de la plateforme bioinformatique GenOuest – Présentation *Contributions of ontologies to life sciences*
novembre 2011 :	École thématique biologie intégrative BioGenOuest – Présentation caractérisation et comparaison fonctionnelles de listes de gènes : apport de la sémantique.
août 2014	article « La bioinformatique avec biopython » dans le hors-série Python de GNU/Linux magazine
décembre 2014 :	Séminaire invité LINA, Nantes.
février 2015 :	Séminaire invité Institut de recherche sur les maladies génétiques Imagine, Paris
mars 2015 :	Séminaire invité ENS Rennes
2014–2015 :	Blog bioinfo-fr.net : articles « Gephi pour la visualisation et l'analyse de graphes » [1], « Gérer les versions de vos fichiers : premiers pas avec git » [2] et « Git : cloner un projet, travailler à plusieurs et créer des branches » [3]

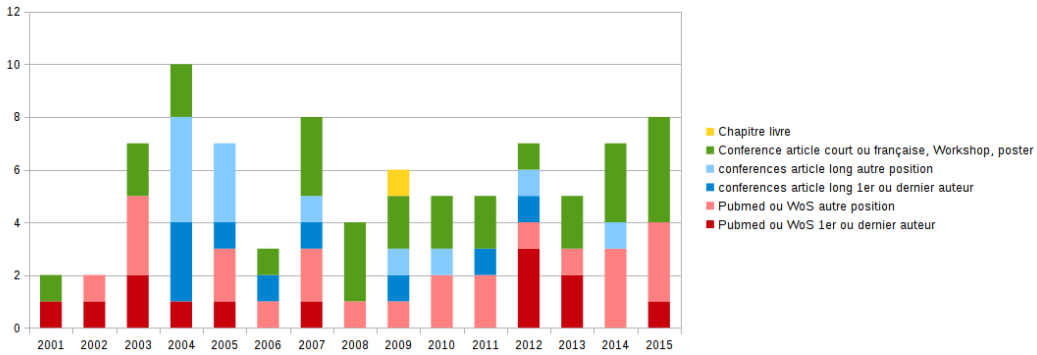

FIG. 2 – *Nombre et type de publications par années*

1. http://bioinfo-fr.net/gephi-pour-la-visualisation-et-lanalyse-de-graphes
2. http://bioinfo-fr.net/git-premiers-pas
3. http://bioinfo-fr.net/git-usage-collaboratif

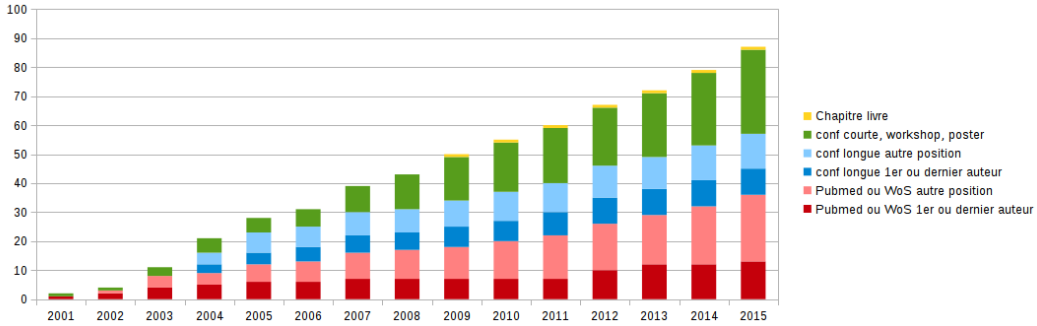

FIG. 3 – *Évolution des publications*

PUBLICATIONS PARUES OU ACCEPTÉES : LISTE EXHAUSTIVE

Bilan :

- H-index : 15
- 18 articles dans des journaux indexés par PubMed ou Web of science
  - 4 en premier auteur
  - 5 en dernier auteur
- 21 articles longs dans des conférences internationales indexés par PubMed ou Web of science avec comité de lecture
  - 6 en premier auteur
  - 1 en dernier auteur

ARTICLES INDEXÉS DANS PUBMED OU WEB OF SCIENCE

[1] Olivier Dameron, Bernard Gibaud, and Xavier Morandi. “Numeric and Symbolic Representation of the Cerebral Cortex Anatomy: Methods and preliminary results”. In: *Surgical and Radiologic Anatomy* 26.3 (2004), pp. 191–197.

[2] Daniel L. Rubin, Olivier Dameron, Yasser Bashir, David Grossman, Parvati Dev, and Mark A. Musen. “Using ontologies linked with geometric models to reason about penetrating injuries”. In: *Artificial Intelligence in Medicine* 37.3 (2006), pp. 167–176.

[3] Olivier Dameron, Mark A. Musen, and Bernard Gibaud. “Using semantic dependencies for consistency management of an ontology of brain-cortex anatomy”. In: *Artificial Intelligence in Medicine* 39.3 (2007), pp. 217–225.

[4] Amrapali Zaveri, Luciana Cofiel, Jatin Shah, Shreyasee Pradhan, Edwin Chan, Olivier Dameron, Ricardo Pietrobon, and Beng Ti Ang. “Achieving High Research Reporting Quality Through the Use of Computational Ontologies”. In: *Neuroinformatics* 8.4 (2010), pp. 261–271.

[5] A Burgun, A Rosier, L Temal, J Jacques, R Messai, L Duchemin, L Deleger, C Grouin, P Van Hille, P Zweigenbaum, R Beuscart, D Delerue, O Dameron, P Mabo, and C Henry. “Decision support in telecardiology: An ontology-based patient-centered approach”. In: *IRBM* 32.3 (2011), pp. 191–194.

[6] Charles Bettembourg, Christian Diot, Anita Burgun, and Olivier Dameron. “GO2PUB: Querying PubMed with Semantic Expansion of Gene Ontology Terms”. In: *Journal of biomedical semantics* 3.1 (2012), p. 7.

[7] Marc Cuggia, Jean-Charles Dufour, Oussama Zekri, Isabelle Gibaud, Cyril Garde, Catherine Bohec, Régis Duvauferrier, Dominique Fieschi, Paolo Besana, Laurent Charlois, Annabel Bourdé, Nicolas Garcelon, Jean-Francois Laurent, Marius Fieschi, and Olivier Dameron. “ASTEC Automatic Selection of clinical Trials based on Eligibility Criteria”. In: *IRBM* (2012).

[8] Olivier Dameron, Charles Bettembourg, and Nolwenn Le Meur. “Measuring the Evolution of Ontology Complexity: the Gene Ontology Case Study”. In: *PLoS ONE* 8.10 (2013), e75993.

[9] Olivier Dameron, Paolo Besana, Oussama Zekri, Annabel Bourdé, Anita Burgun, and Marc Cuggia. “OWL Model of Clinical Trial Eligibility Criteria Compatible with Partially-known Information”. In: *Journal of Biomedical Semantics* 4.1 (2013).

[10] Jean-Francois Ethier, Olivier Dameron, Vasa Curcin, Mark M. McGilchrist, Robert A. Verheij, Theodoros N. Arvanitis, Adel Taweel, Brendan C. Delaney, and Anita Burgun. “A Unified Structural/Terminological Framework based on LexEVS: application to TRANSFoRm”. In: *Journal of the American Medical Informatics Association* 20.5 (2013), pp. 986–994.

[11] Charles Bettembourg, Christian Diot, and Olivier Dameron. “Semantic particularity measure for functional characterization of gene sets using Gene Ontology”. In: *PLoS ONE* 9.1 (2014), e86525.

[12] Gautier Defossez, Alexandre Rollet, Olivier Dameron, and Pierre Ingrand. “Temporal representation of care trajectories of cancer patients using data from a regional information system: an application in breast cancer”. In: *BMC Medical Informatics and Decision Making* 14.1 (2014), p. 24.

[13] Frederic Herault, Annie Vincent, Olivier Dameron, Pascale Le Roy, Pierre Cherel, and Marie Damon. “The longissimus and semimembranosus muscles display marked differences in their gene expression profiles in pig”. In: *PloS one* 9.5 (2014), e96491.

[14] Sylvain Prigent, Guillaume Collet, Simon M Dittami, Ludovic Delage, Floriane Ethis de Corny, Olivier Dameron, Damien Eveillard, Sven Thiele, Jeanne Cambefort, Catherine Boyen, Anne Siegel, and Thierry Tonon. “The genome-scale metabolic network of Ectocarpus siliculosus (EctoGEM): a resource to study brown algal physiology and beyond”. In: *The Plant journal : for cell and molecular biology* 80.2 (2014), pp. 367–381.

[15] Charles Bettembourg, Christian Diot, and Olivier Dameron. “Optimal Threshold Determination for Interpreting Semantic Similarity and Particularity: Application to the Comparison of Gene Sets and Metabolic Pathways Using GO and ChEBI”. In: *PloS one* 10.7 (2015), e0133579.

[16] Philippe Finet, Régine Le Bouquin-Jeannès, Olivier Dameron, and Bernard Gibaud. “Review of current telemedicine applications for chronic diseases: Toward a more integrated system?” In: *IRBM* (2015). In press.

[17] Andrej Machno, Pierre Jannin, Olivier Dameron, Werner Korb, Gerik Scheuermann, and Jürgen Meixensberger. “Ontology for assessment studies of human-computer-interaction in surgery”. In: *Artificial intelligence in medicine* 63.2 (2015), pp. 73–84.

[18] Arnaud Rosier, Philippe Mabo, Lynda Temal, Pascal Van Hille, Olivier Dameron, Louise Deléger, Cyril Grouin, Pierre Zweigenbaum, Julie Jacques, Emmanuel Chazard, Laure Laporte, Christine Henry, and Anita Burgun. “Personalized and automated remote monitoring of atrial fibrillation”. In: *Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology* (2015). In press.

Conférences internationales indexées dans PubMed ou Web of science

[1] Olivier Dameron, Bernard Gibaud, and Xavier Morandi. “Numeric and Symbolic Knowledge Representation of Cortex Anatomy Using Web Technologies”. In: *Artificial Intelligence Medicine, 8th Conference on AI in Medicine in Europe, AIME 2001, Cascais, Portugal, July 1-4, 2001, Proceedings*. Ed. by Silvana Quaglini, Pedro Barahona, and Steen Andreassen. Vol. 2101. Lecture Notes in Computer Science. Springer, 2001, pp. 359–68. ISBN: 3-540-42294-3.

[2] Bernard Gibaud, Olivier Dameron, and Xavier Morandi. “Representation and sharing of numeric and symbolic knowledge about brain cortex anatomy using web technology”. In: *Computer Assisted Radiology and Surgery 2001*. Ed. by HU Lemke, MW Vannier, K Inamura, AG Farman, and K Doi. Elsevier, 2001, pp. 356–361.

[3] Olivier Dameron, Bernard Gibaud, Anita Burgun, and Xavier Morandi. “Towards a sharable numeric and symbolic knowledge base on cerebral cortex anatomy: lessons from a prototype”. In: *American Medical Informatics Association AMIA*. 2002, pp. 185–189.

[4] Olivier Dameron, Anita Burgun, Xavier Morandi, and Bernard Gibaud. “Modelling dependencies between relations to insure consistency of a cerebral cortex anatomy knowledge base”. In: *Studies in Health technology and informatics*. 2003, pp. 403–408.

[5] Bernard Gibaud, Olivier Dameron, and Xavier Morandi. “Re-use of a multi-purpose knowledge corpus on cortex anatomy for educational purposes”. In: *Studies in Health technology and informatics*. 2003, pp. 439–444.

[6] Christine Golbreich, Olivier Dameron, Bernard Gibaud, and Anita Burgun. “How to represent ontologies in view of a Medical Semantic Web ?” In: *AIME 03 Conference Proceedings*. 2003, pp. 51–60.

[7] Christine Golbreich, Olivier Dameron, Bernard Gibaud, and Anita Burgun. “Web ontology language requirements w.r.t expressiveness of taxononomy and axioms in medecine”. In: *International Semantic Web Conference ISWC03 proceedings*. Vol. 2870. Lecture Notes in Computer Science. Springer, 2003.

[8] Olivier Dameron, Natalya F. Noy, Holger Knublauch, and Mark A. Musen. “Accessing and Manipulating Ontologies Using Web Services”. In: *Proceeding of the Third International Semantic Web Conference (ISWC2004), Semantic Web Services workshop*. 2004.

[9] Olivier Dameron, Daniel L. Rubin, and Mark A. Musen. “Challenges in Converting Frame-Based Ontology into OWL: the Foundational Model of Anatomy Case-Study”. In: *American Medical Informatics Association Conference AMIA05*. 2005, pp. 181–185.

[10] Bernard Gibaud, Olivier Dameron, Éric Poiseau, and Pierre Jannin. “Implementation of atlas-matching capabilities using Web Services technology: lessons learned from the development of a demonstrator”. In: *Computer Assisted Radiology and Surgery 2005*. 2005.

[11] Daniel L. Rubin, Olivier Dameron, and Mark A. Musen. “Use of Description Logic Classification to Reason about Consequences of Penetrating Injuries”. In: *American Medical Informatics Association Conference AMIA05*. 2005, pp. 649–653.

[12] Gwenaëlle Marquet, Olivier Dameron, Stephan Saikali, Jean Mosser, and Anita Burgun. “Grading glioma tumors using OWL-DL and NCI Thesaurus”. In: *Proceedings of the American Medical Informatics Association Conference AMIA'07*. 2007, pp. 508–512.

[13] Elena Beisswanger, Vivian Lee, Jung-Jae Kim, Dietrich Rebholz-Schuhmann, Andrea Splendiani, Olivier Dameron, Stefan Schulz, and Udo Hahn. “Gene Regulation Ontology (GRO): design principles and use cases”. In: *Studies in Health technology and informatics - Proceedings of the Medical Informatics in Europe conference (MIE'08)*. Vol. 136. 2008, pp. 9–14.

[14] Cyril Grouin, Arnaud Rosier, Olivier Dameron, and Pierre Zweigenbaum. “Testing tactics to localize de-identification”. In: *Studies in health technology and informatics* 150 (2009), pp. 735–739.

[15] Anita Burgun, Lynda Temal, Arnaud Rosier, Olivier Dameron, Philippe Mabo, Pierre Zweigenbaum, Regis Beuscart, David Delerue, and Henry Christine. “Integrating clinical data with information transmitted by implantable cardiac defibrillators to support medical decision in telecardiology: the application ontology of the AKENATON project”. In: *Proceedings of the American Medical Informatics Association Conference AMIA*. 2010, p. 992.

[16] Lynda Temal, Arnaud Rosier, Olivier Dameron, and Anita Burgun. “Mapping BFO and DOLCE”. In: *Studies in health technology and informatics* 160 (2010), pp. 1065–1069.

[17] Marc Cuggia, Jean-Charles Dufour, Paolo Besana, Olivier Dameron, Regis Duvauferrier, Dominique Fieschi, Catherine Bohec, Annabel Bourdé, Laurent Charlois, Cyril Garde, Isabelle Gibaud, Jean-Francois Laurent, Oussama Zekri, and Marius Fieschi. “ASTEC: A System for Automatic Selection of Clinical Trials”. In: *Proceedings of the American Medical Informatics Association Conference AMIA*. 2011, p. 1729.

[18] Olivier Dameron, Pascal van Hille, Lynda Temal, Arnaud Rosier, Louise Deléger, Cyril Grouin, Pierre Zweigenbaum, and Anita Burgun. “Comparison of OWL and SWRL-Based Ontology Modeling Strategies for the Determination of Pacemaker Alerts Severity”. In: *Proceedings of the American Medical Informatics Association Conference AMIA*. 2011, p. 284.

[19] Cyril Grouin, Louise Deléger, Arnaud Rosier, Lynda Temal, Olivier Dameron, Pascal van Hille, Anita Burgun, and Pierre Zweigenbaum. “Automatic computation of CHA2DS2-VASc score: Information extraction from clinical texts for thromboembolism risk assessment”. In: *Proceedings of the American Medical Informatics Association Conference AMIA*. 2011, pp. 501–510.

[20] Pascal van Hille, Julie Jacques, Julien Taillard, Arnaud Rosier, David Delerue, Anita Burgun, and Olivier Dameron. "Comparing Drools and Ontology-based reasoning approaches for telecardiology decision support". In: *Studies in health technology and informatics* 180 (2012), pp. 300–304.

[21] A Machno, P Jannin, O Dameron, W Korb, G Scheuermann, and J Meixensberger. "Analysis of MCI evaluation studies in surgery". In: *Proceedings of the Computer Assisted Radiology and Surgery conference CARS2012*. In press. 2012.

Conférences internationales avec comité non indexées

[1] Christian Barillot, Romain Valabregue, Jean-Pierre Matsumoto, Florent Aubry, Habib Benali, Yann Cointepas, Olivier Dameron, Michel Dojat, E. Duchesnay, Bernard Gibaud, Serge Kinkingnéhun, Dimitri Papadopoulos, Mélanie Pellegrini-Issac, and Éric Simon. "Neurobase: Management of Distributed and Heterogeneous Information Sources in Neuroimaging". In: *MICCAI 2004 Conference*. Ed. by M. Dojat and B. Gibaud. 2004, pp. 85–94.

[2] Olivier Dameron. "JOT: a Scripting Environment for creating and managing ontologies". In: *7th International Protégé Conference*. 2004.

[3] Olivier Dameron. "Using the JOT plugin for reasoning with Protégé". In: *Workshop on Protégé and Reasoning – 7th International Protégé Conference*. 2004.

[4] Olivier Dameron, Bernard Gibaud, and Mark Musen. "Using semantic dependencies for consistency management of an ontology of brain-cortex anatomy". In: *First International Workshop on Formal Biomedical Knowledge Representation KRMED04*. 2004, pp. 30–38.

[5] Olivier Dameron and Mark A. Musen. "Accessing and manipulating Life-Sciences Ontologies Using Web Services". In: *W3C Workshop on Semantic Web for Life Sciences*. 2004.

[6] Bernard Gibaud, Michel Dojat, Habib Benali, Olivier Dameron, Jean-Pierre Matsumoto, Mélanie Pellegrini-Issac, Romain Valabregue, and Christian Barillot. "Toward an Ontology for Sharing Neuroimaging Data and Processing Methods: Experience Learned from the Development of a Demonstrator". In: *MICCAI 2004 Conference*. Ed. by M. Dojat and B. Gibaud. 2004, pp. 15–23.

[7] Holger Knublauch, Olivier Dameron, and Mark Musen. "Weaving the Biomedical Semantic Web with the Protégé OWL Plugin". In: *First International Workshop on Formal Biomedical Knowledge Representation KRMED04*. 2004, pp. 39–47.

[8] D.L. Rubin, O. Dameron, Y. Bashir, D. Grossman, P. Dev, and M.A. Musen. "Using ontologies linked with geometric models to reason about penetrating injuries". In: *Intelligent Data Analysis in Medicine and Pharmacology IDAMAP04*. 2004.

[9] Olivier Dameron. "Keeping modular and platform-independent software up-to-date: benefits from the Semantic Web". In: *8th International Protégé Conference*. 2005.

[10] Christine Golbreich, Olivier Bierlaire, Olivier Dameron, and Bernard Gibaud. "Use Case: Ontology with Rules for identifying brain anatomical structures". In: *W3C Workshop on Rule Languages for Interoperability*. 2005.

[11] Christine Golbreich, Olivier Bierlaire, Olivier Dameron, and Bernard Gibaud. "What reasoning support for ontology and rules? the brain anatomy case study". In: *8th International Protégé Conference*. 2005.

[12] Daniel L. Rubin, Olivier Dameron, and Mark A. Musen. "Using OWL and Description Logics Based Classification for Reasoning in Biomedical Applications". In: *8th International Protégé Conference*. 2005.

[13] Olivier Dameron, Élodie Roques, Daniel L. Rubin, Gwenaëlle Marquet, and Anita Burgun. "Grading lung tumors using OWL-DL based reasoning". In: *9th International Protégé Conference*. 2006.

[14] Julie Chabalier, Olivier Dameron, and Anita Burgun. "Integrating and querying disease and pathway ontologies: building an OWL model and using RDFS queries". In: *Bio-Ontologies Special Interest Group, Intelligent Systems for Molecular Biology conference (ISMB'07)*. 2007.

[15] Julie Chabalier, Olivier Dameron, and Anita Burgun. "Integrating disease and pathway ontologies". In: *Proceedings of the ISMB conference, Poster Session*. 2007.

[16] Olivier Dameron and Julie Chabalier. "Automatic generation of consistency constraints for an OWL representation of the FMA". In: *10th International Protégé Conference*. 2007.

[17] Andrea Splendiani, Elena Beisswanger, Jung-Jae Kim, Vivian Lee, Olivier Dameron, and Dietrich Rebholz-Schuhmann. "Bio-Ontologies in the context of the BOOTStrep project". In: *Proceedings of the Bio-Ontologies SIG Workshop ISMB, (poster)*. 2007.

[18] Olivier Dameron and Julie Chabalier. "Bio-ontologies Tutorial". In: *Proceedings of the Data Integration In Life Science conference DILS*. Ed. by A. Bairoch, S. Cohen-Boulakia, and Froidevaux C. Vol. 5109. LNBI. 2008, p. 208.

[19] Olivier Dameron, Charles Bettembourg, and Léa Joret. "Quantitative cross-species comparison of GO annotations: advantages and limitations of semantic similarity measure". In: *11th International Protégé Conference*. 2009.

[20] Lynda Temal, Arnaud Rosier, Olivier Dameron, and Anita Burgun. "Modeling cardiac rhythm and heart rate using BFO and DOLCE". In: *International Conference for Biomedical Ontologies*. 2009.

[21] Anita Burgun, Arnaud Rosier, Lynda Temal, Olivier Dameron, Philippe Mabo, Pierre Zweigenbaum, Régis Beuscart, David Delerue, and Christine Henry. "Supporting medical decision in telecardiology: a patient-centered ontology-based approach". In: *Medinfo 2010*. In press. 2010.

[22] Nicolas Lebreton, Christophe Blanchet, Daniela Barreiro Claro, Julie Chabalier, Anita Burgun, and Olivier Dameron. "Verification of parameters semantic compatibility for semi-automatic Web service composition: a generic case study". In: *12th International Conference on Information Integration and Web-based Applications and Services (iiWAS2010)*. 2010, pp. 845–848.

[23] Olivier Dameron, Paolo Besana, Oussama Zekri, Annabel Bourdé, Anita Burgun, and Marc Cuggia. "OWL Model of Clinical Trial Eligibility Criteria Compatible With Partially-known Information". In: *Proceedings of the Semantic Web for Life Sciences workshop SWAT4LS2012*. 2012.

[24] Wiktoria Golik, Olivier Dameron, Jérôme Bugeon, Alice Fatet, Isabelle Hue, Catherine Hurtaud, Matthieu Reichstadt, Marie-Christine Salaün, Jean Vernet, Léa Joret, Frédéric Papazian, Claire Nédellec, and Pierre-Yves Le Bail. "ATOL: the multi-species livestock trait ontology". In: *Proceedings of the 6th Metadata and Semantics Research Conference MTSR*. 2012.

[25] Anthony Bretaudeau, Olivier Dameron, Fabrice Legeai, and Yvan Rahbé. "AphidAtlas : avancées récentes". In: *Proceedings of BAPOA 2013 MOP. INRA, CIRAD Lavalette Campus Montpellier, France*. Ed. by M. Uzest. [In French]. 2013.

[26] Isabelle Hue, Jérôme Bugeon, Olivier Dameron, Alice Fatet, Catherine Hurtaud, Léa Joret, Marie-Christine Meunier-Salaün, Claire Nédellec, Matthieu Reichstadt, Jean Vernet, and Pierre-Yves Le Bail. "ATOL and EOL ontologies, steps towards embryonic phenotypes shared worldwide?" In: *Proceedings of the 4th Mammalian Embryo Genomics Meeting, October 2013, Quebec City*. Vol. 149. Animal Reproduction Science 1–2. 2014, p. 99.

[27] François Moreews, Yvan Le Bras, Olivier Dameron, Cyril Monjeaud, and Olivier Collin. "Integrating GALAXY workflows in a metadata management environment". In: *Galaxy Community Conference GCC2014, Proceedings*. 2014. URL: `https://wiki.galaxyproject.org/Events/GCC2014/Abstracts/Posters#P28:_Integrating_GALAXY_workflows_in_a_metadata_management_environment`.

Articles de journaux nationaux avec comité

[1] JJ Levrel, B Carsin-Nicol, C Ouail-Tabourel, E Chabert, P Darnault, B Gibaud, O Dameron, and X Morandi. "Electronic imaging with photo-realistic rendering for neuroanatomy teaching: methods and preliminary results". In: *Journal of Neuroradiology* (2002).

[2] V Bertaud, I Belhadj, O Dameron, N Garcelon, L Hendaoui, F Marin, and R Duvauferrier. "L'informatisation du signe radiologique". In: *Journal de Radiologie* 88.1 (2007), pp. 27–37.

[3] Philippe Finet, Régine Le Bouquin-Jeannès, and Olivier Dameron. "La télémédecine dans la prise en charge des maladies chroniques [in French]". In: *Techniques Hospitalières* 740 (2013).

[4] Pierre-Yves Le Bail, Jérôme Bugeon, Olivier Dameron, Alice Fatet, Wiktoria Golik, Jean-François Hocquette, Catherine Hurtaud, Isabelle Hue, Catherine Jondreville, Léa Joret, Marie-Christine Meunier-Salaün, Jean Vernet, Claire Nedellec, Matthieu Reichstadt, and Philippe Chemineau. "Un langage de référence pour le phénotypage des animaux d'élevage : l'ontologie ATOL". In: *Production Animale* 27.3 (2014), pp. 195–208.

Conférences nationales avec comité non indexées

[1] Olivier Dameron, Bernard Gibaud, and Xavier Morandi. "Représentation de connaissances numériques et symboliques sur l'anatomie du cortex cérébral par des technologies du web". In: *Forum du Jeune Chercheur Compiègne*. 2001.

[2] Olivier Dameron, Anita Burgun, Xavier Morandi, and Bernard Gibaud. "Ontologie stratifiée de l'anatomie du cortex cérébral : application au maintien de la cohérence". In: *Journée Web Sémantique, Rennes*. 2003.

[3] Christine Golbreich, Olivier Dameron, Bernard Gibaud, and Anita Burgun. "Comment représenter les ontologies pour tendre vers un Web Sémantique Médical ?" In: *Journées Fran caises de la Toile*. 2003.

[4] Julie Chabalier, Gwenaëlle Marquet, Olivier Dameron, and Anita Burgun. "Enrichissement de la hiérarchie KEGG par l'exploitation de Gene Ontology". In: *Workshop OGSB, JOBIM'06*. 2006.

[5] Julie Chabalier, Olivier Dameron, and Anita Burgun. "Using knowledge about pathways as an organizing principle for disease ontologies". In: *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM'07)*. 2007.

[6] Nicolas Lebreton, Olivier Dameron, Christophe Blanchet, and Julie Chabalier. "Utilisation d'ontologies de tâches et de domaine pour la composition semi-automatique de services Web bioinformatiques". In: *Proceedings of the Journées Ouvertes de Biologie, Informatique et Mathématiques (Jobim 2008)*. 2008.

[7] Élodie Roques, Julie Chabalier, and Olivier Dameron. "Enrichissement sémantique de patrons syntaxiques pour l'amélioration du mapping entre voies métaboliques et processus biologiques". In: *Proceedings of the Journées Ouvertes de Biologie, Informatique et mathématiques (Jobim 2008)*. 2008.

[8] Cyril Grouin, Arnaud Rosier, Olivier Dameron, and Pierre Zweigenbaum. "Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers". In: *Journées Francophones d'Informatique Médicale*. 2009.

[9] Nicolas Lebreton, Christophe Blanchet, Julie Chabalier, and Olivier Dameron. "Utilisation d'ontologies de tâches et de domaine pour la composition semi-automatique de services Web bioinformatiques". In: *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2009)*. 2009.

[10] Charles Bettembourg, Christian Diot, and Olivier Dameron. "Cross-Species Metabolic Pathways Comparison: Focus on Mouse, Human and Chicken Lipid Metabolism". In: *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2011)*. In press. 2011.

[11] Charles Bettembourg, Christian Diot, Anita Burgun, and Olivier Dameron. "GO2PUB: Querying PubMed with Semantic Expansion of Gene Ontology Terms". In: *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2012)*. 2012.

[12] Alexandre Rollet, Gautier Defossez, Olivier Dameron, Poitou-Charentes CoRIM, Poitou-Charentes CRISAP, and Pierre Ingrand. "Développement et évaluation d'un algorithme de représentation des parcours de soins de patientes atteintes d'un cancer du sein à partir des données d'un système d'information régional". In: *Proceedings of the conference Évaluation Management Organisation Information Santé (EMOIS2013, Nancy, France)*. [In French, short abstract]. 2013.

[13] Charles Bettembourg, Olivier Dameron, Anthony Bretaudeau, and Fabrice Legeai. "Intégration et interrogation de réseaux de régulation génomique et post-génomique". In: *Proceedings of the INOVIVE workshop (INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement), conférence IC (Ingénierie des Connaissances) PFIA*. 2015.

[14] Jean Coquet, Geoffroy Andrieux, Jacques Nicolas, Olivier Dameron, and Nathalie Theret. “Analysis of TGF-beta signalization pathway thanks to topological and Semantic Web methods”. In: *Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM 2015), poster session*. 2015.

[15] Philippe Finet, Bernard Gibaud, Olivier Dameron, and Régine Le Bouquin-Jeannès. “Interopérabilité d’un système de capteurs en télémédecine”. In: *Proceedings of the Journées d’étude sur la Télésanté, UTC Compiègne*. 2015.

[16] Yann Rivault, Olivier Dameron, and Nolwenn Le Meur. “Une infrastructure générique basée sur les apports du Web Sémantique pour l’analyse des bases médico-administratives”. In: *Proceedings of the IN-OVIVE workshop (INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l’Environnement), conférence IC (Ingénierie des Connaissances) PFIA*. 2015.

[17] Yann Rivault, Olivier Dameron, and Nolwenn Le Meur. “La gestion de données médico-administratives grâce aux outils du Web Sémantique”. In: *Proceedings of the conference Évaluation Management Organisation Information Santé (EMOIS2016, Dijon, France)*. [In French, short abstract]. 2016.

Divers

Voile :	Moniteur fédéral de voile (catamaran). Participation au challenge CNRS en 2010, 2011, 2012, 2013, 2014 et 2015.
Associations :	Président d’une crèche parentale (février 2010 – juin 2011) ; Trésorier de l’« association voile recherche-enseignement Rennes » (depuis 2012).

18 décembre 2015