Semi-automated hypothesis evaluation using semantic technologies

by

Alison Victoria Callahan

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Biology

Carleton University
Ottawa, Ontario

Thesis Abstract

In today's age of "big data" and "omics" research, biologists face two unique challenges - sharing their results with the larger community in an interpretable and reusable format and integrating their experimental data and findings with the prevailing hypotheses that govern their field. Publicly funded biological data curation and warehousing centers have emerged to address the former, but the challenge remains of sifting out relevant information from these resources and integrating it in a scalable way towards assessing biological hypotheses, and in disseminating the results of this process. To address these challenges, I have developed, implemented and evaluated a semi-automated system for biological hypothesis evaluation that uses semantic technologies to reason over existing experimental data and knowledge. Chapter 1 presents the motivation, driving hypothesis and objectives for this doctoral thesis, as well as a brief review of the Semantic Web and automated systems for hypothesis formulation and evaluation. In Chapter 2 I present HyQue, a Semantic Web tool for evaluating scientific hypotheses, including the system architecture and a prototype implementation for evaluating hypotheses about yeast metabolism. In Chapter 3, I describe efforts to publish and integrate biological data on the Semantic Web through the Bio2RDF project, a key data source for HyQue that enables browsing, querying and downloading over 3 billion statements from more than 25 life sciences databases. In Chapter 4 I describe the ovopub, a linked data model for capturing provenance on the Semantic Web, as well as its implementation and application to Bio2RDF data. The ovopub provides a simple model for describing basic elements of linked data provenance, and enables provenance-based querying and filtering over biological linked data. In Chapter 5 I describe the application of HyQue to evaluating

hypotheses about the role of *C. elegans* genes in aging. HyQue correctly identified known lifespan-related genes, as well as 24 candidate aging-related genes by retrieving and evaluating domain-specific evidence from multiple sources. Chapter 6 summarizes the contributions of this thesis and proposes future work.

Acknowledgements

I would like to thank my supervisor, Dr. Michel Dumontier, for teaching me about bioinformatics and programming from the time that I was an undergraduate, and for motivating me to pursue research outside of the classroom. My perspective on learning and scientific research was fundamentally changed by our first summer of work together. Since that time you have continually guided and supported my efforts to develop new skills, and have been unwavering in your willingness to critically evaluate and provide feedback on my research, all the while encouraging me to work independently.

In addition to the research opportunities I have been afforded during my doctoral work, I have had the great privilege of sharing our lab with bright and wonderful individuals. I thank Natalia Villanueva-Rosales for being an excellent role model, and for always bringing a warm and positive energy to the room. Special thanks to Glen Newton for your invaluable friendship and support over the years, and for generously sharing your extensive technical knowledge and critical thinking skills.

My parents, Joe and Nancy, have been instrumental in my development as a scientist and a person. Thank you for fostering my interest in exploring the natural world, giving me just the right amount of encouragement to pursue an education in the sciences, and for everything you've done to facilitate and support a life of learning. I am grateful to my brother Chris for always being interested, and for giving me a perspective on life and work I would never have found on my own.

Infinite thanks to José Miguel Cruz Toledo for motivating me every day to work hard, for setting an example with your ever-inquiring mind and true passion for science, and for giving me courage each step of the way.

Preface

Co-authorship statement for Chapters 2, 3, and 4

Of the four data chapters that comprise the bulk of this thesis, three exist in a previously published form. Chapter 2 was published as a chapter in the LNCS book *The Semantic Web: Research and Applications*[*]. Michel Dumontier and I jointly conceived of the research idea. Under the supervision of Michel Dumontier, I developed the HyQue framework, applied it to biological hypotheses, and analyzed the results. Michel Dumontier and I wrote the manuscript. The work presented in Chapter 2 is an extension of work described in:

> Callahan, Alison, Michel Dumontier and Nigam H. Shah. 2011. HyQue: evaluating hypotheses using Semantic Web technologies. *Journal of Biomedical Semantics* 2 (Suppl 2): S3.

Chapter 2 differs significantly from this work in two respects: (1) The version of HyQue described in the 2011 JBMS article ('v1') was implemented using the Hypertext Preprocessor (PHP) scripting language, while the extended version of HyQue described in Chapter 2 ('v2') is implemented in Java. (2) HyQue 'v1' does not use the SPARQL Inferencing Notation (SPIN) while HyQue 'v2' rules and functions are entirely defined using SPIN, with the advantage that they are themselves described in RDF and executed using the SPIN API. Chapter 3 was published as a chapter in the Lecture Notes in Computer Science (LNCS) book *The Semantic Web: Semantics and Big Data*[*]. Michel Dumontier, José Cruz Toledo and I jointly conceived of the research idea. Under the supervision of Michel Dumontier, José Cruz Toledo and I developed and executed

[*] The complete bibliographic information for published chapters 2 and 3 are provided at the beginning of each chapter.

updated Bio2RDF scripts, generated Bio2RDF ontology-SIO mappings, and developed and analyzed queries over the resulting datasets. Michel Dumontier, José Cruz Toledo and I composed the manuscript. Peter Ansell had previously developed the Bio2RDF Webapp and contributed to its configuration for Bio2RDF Release 2 and its description in the manuscript. A previous version of Chapter 4 is deposited in the arXiv repository as:

> Callahan, Alison and Michel Dumontier. 2013. Ovopub: Modular data publication with minimal provenance. arXiv:1305.6800 [cs.DL].

The work presented in Chapter 4 is an extension of the work described in the arXiv manuscript that includes an implementation of ovopubs for the iRefIndex Bio2RDF dataset and demonstrates queries over ovopub-specific and iRefindex-sourced provenance. Michel Dumontier and I jointly conceived of the research idea. Under the supervision of Michel Dumontier, I developed the ovopub model, used it to generate ovopubs from Bio2RDF linked data, and analyzed the utility of the model. Michel Dumontier and I wrote the manuscript.

Contribution of published works to thesis

For each of the published works included as chapters in this thesis, I describe the contribution of the work towards achieving the objectives of my doctoral research in a prelude section entitled ‘Contribution to thesis’ that follows the abstract accompanying each chapter. This purpose of this section is to situate the work in the broader context of the thesis.

Table of Contents

List of Tables

List of Figures

List of Appendices

1 Chapter: General Introduction

"The challenge of the Semantic Web … is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web"

(Berners-Lee, 2001)

"… creating technologies that represent and interpret multiple, diverse data sources and that support collaborative scientific interpretation of these sources is critical"

(Altman et al., 1999)

1.1 Motivation

One of the central tenets of experimentation in the life sciences is that the formulation and testing of hypotheses will lead to an improved understanding of biological systems that cannot be directly observed due to their complexity and size [1]. Valid hypotheses build on the results of prior experiments and can be tested through new experimentation. However, with hundreds of thousands of relevant articles published each year in scientific journals [2] and terabytes of related data, scientists are overwhelmed and unable to sift through and collate all essential facts to support or dispute a hypothesis. In addition to *testing* hypotheses through new experimentation, significant work is required to *evaluate* hypotheses in the context of existing data. Indeed, biologists perceive that the predominant challenge in research is to "locate, integrate and access" the vast amounts of biological data resulting from small- and large-scale experiments [3]. Improving scientists' ability to effectively integrate information has significant consequences for the scientific enterprise in terms of saving time, preventing unnecessary duplication of research efforts, lowering costs and increasing productivity [4]. Approaches that truly encompass the promise of 'Big Data' by providing services for processing data at scale and policies to govern these services, *in addition* to the data itself, have the potential to foster scientific collaboration and accelerate discovery, likely in ways that we cannot currently envision [5].

E-Science [6-8] has much to gain from a system for automated hypothesis evaluation that brings question-answering and the distributed use and analysis of scientific data up to the scale of current data production in the natural sciences. Novel approaches directed towards the information integration challenge could significantly

change the way that scientists interact and publish scientific results [9], enabling a whole new paradigm around using that information to develop effective experiments that improve our overall understanding. Systems for scientists that facilitate reproducibility, generate machine understandable results, and enable data retrieval informed by the type of analysis executed to generate that data, are necessary to advance the discovery level and value of natural sciences research in this information-centric age. Motivated by this potential, my doctoral research aims to test the following hypothesis within the framework of two broad objectives.

1.2 Hypothesis

Formal machine-understandable models for hypotheses, hypothesis evaluation criteria, scientific knowledge and data enable the semi-automated evaluation of biological hypotheses over existing knowledge and experimental data.

1.3 Research objectives

1. Develop a framework for formulating and evaluating scientific hypotheses using formally represented data and knowledge
2. Implement this framework to evaluate biological hypotheses

1.4 Thesis outline

In the remaining sections of this chapter, I briefly review Semantic Web technologies and automated systems for hypothesis formulation and evaluation. In Chapter 2 I present the architecture and implementation of HyQue, a Semantic Web tool for evaluating scientific hypotheses, and its prototype biological application. In Chapter 3 I describe the Bio2RDF project, a key resource for HyQue that enables browsing, querying and downloading over 3 billion statements from more than 25 life sciences databases that have been structured

to enable data integration and reasoning using the Semanticscience Integrated Ontology. In Chapter 4 I describe the ovopub, a linked data model for capturing provenance on the Semantic Web and its application to the iRefindex database of protein-protein interactions (PPI), including domain specific descriptions of PPI experimental provenance that are important in assessing the strength of evidence associated with a PPI. In Chapter 5 I describe the application of HyQue for evaluating hypotheses about the role of *C. elegans* genes in aging, and demonstrate that HyQue can identify known aging- and longevity-associated genes as well as quantify experimental support for 24 candidate genes. Chapter 6 summarizes the contributions of this thesis and proposes future work.

1.5 Background

1.5.1 The Semantic Web: Making the Web machine-understandable

The HyperText Markup Language (HTML), used to create virtually all Web pages we visit, structures online documents primarily in terms of layout and is used for formatting how a page looks as well as linking pages to each other on the Web. In concert with Web technologies such as Cascading Style Sheets (CSS), JavaScript and the AJAX paradigm, this simple language has remained the foundation for building complex, beautiful and significant online experiences (Figure 1). However, these experiences are limited to what can be processed and understood by humans. HTML does not describe the content within or meaning of the links between pages in any machine-understandable way. We must rely on our own interpretation to determine the context of links between Web pages, and how their content is (or is not) related. Computers are left mostly "in the dark" as to the meaning of Web pages' content and the links between them, unless this content is further analyzed using approaches such as natural language processing. HTML5, the most recent

release of HTML, includes more meaning-focused tags (dubbed ‘semantic elements’) such as <aside>, <figure>, <section>, and <summary>, in addition to the long-standing <table> and <img>, but they lack a machine-interpretable definition and what lies between these tags, as well as information about their creator and source remains, from the perspective of the computer, simply a sequence of characters.

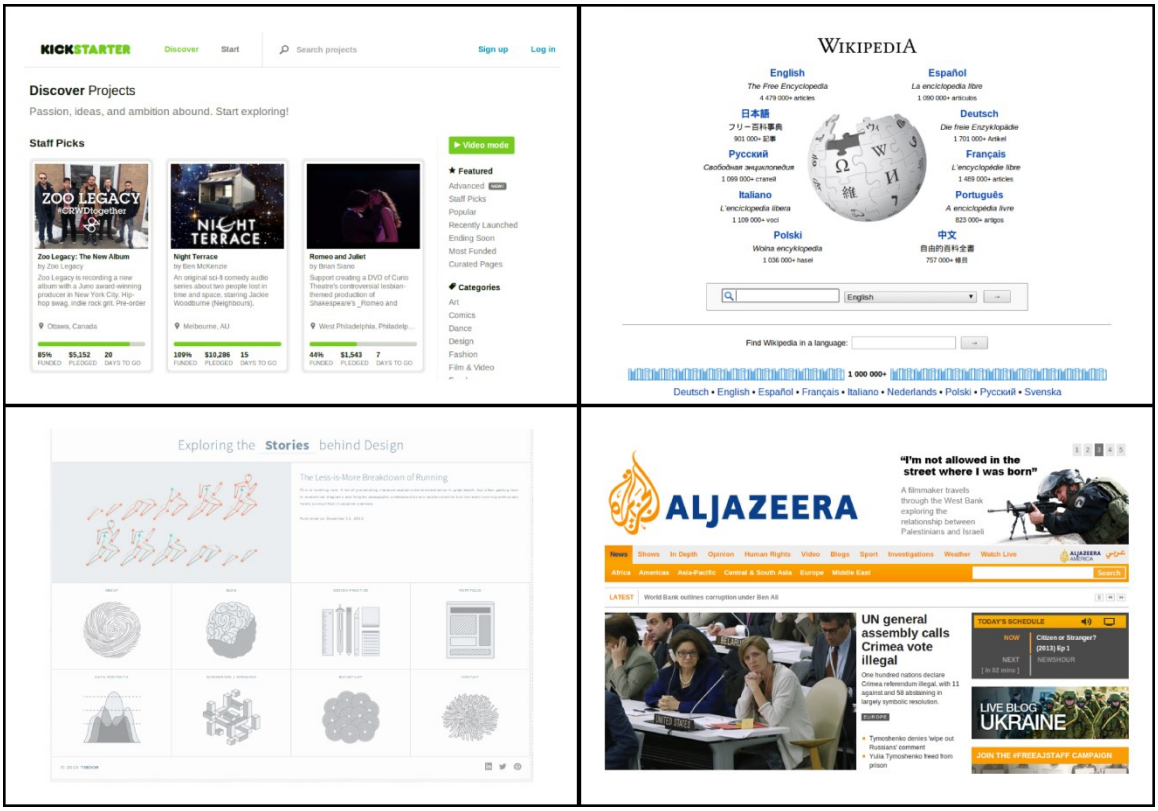


Figure 1 HTML based websites are an everyday part of people’s lives, providing news and information, beautiful designs, and human connection, but leave computers mostly “in the dark”. Clockwise from top left: kickstarter.com, wikipedia.com, aljazeera.com and yeedor.com.

Motivated by this problem, and envisioning a world where software agents on the Web could process and act on machine-understandable data, Tim Berners-Lee first conceived of the Semantic Web [10] to facilitate knowledge representation, information sharing and data integration in a distributed, decentralized manner, through a standard set of machine-understandable languages and protocols with a well-defined *syntax* and *semantics*.

Vocabularies that define concepts for structuring and linking data can themselves be published on the Semantic Web, and the links or relationships between them also described, thereby providing both humans and computers with the ability to interpret their meaning. The technologies at the core of the Semantic Web form a 'stack' (Figure 2) with the most basic requirements (unique and resolvable identifiers) at its foundation and more complex technologies for representing knowledge and querying facilitating web-based logic, leading eventually to support for trusted machine-enabled communications. This stack also represents the progression of Semantic Web technologies, gaining more functionality through time.

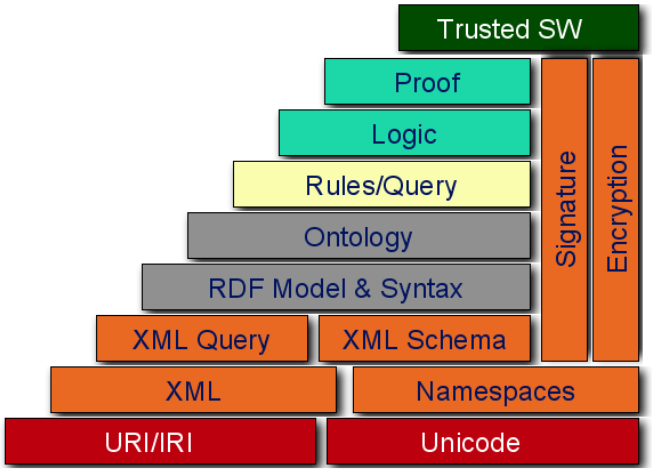


Figure 2 The Semantic Web technology stack [11].

1.5.1.1 RDF, Linked Data and SPARQL

The Resource Description Framework (RDF) is a simple but powerful data model for the Semantic Web that describes facts as collections of subject-predicate-object triples, each element of which can be typed by taxonomically organized vocabularies. Using RDF triples, it is possible to describe the relationships between and properties of entities. Consider the statement 'Alison has age 29'. Here, the subject is 'Alison', the predicate is

'has age' and the object is '29'. 'Alison' could be assigned a resource with a unique identifier, *e.g.* 'http://example.org/people/Alison'. In RDF (using the Turtle or TTL syntax), this statement could be expressed as:

```
@prefix people: <http://example.org/people/> .
@prefix quality: <http://example.org/quality/> .
people:Alison quality:has_age "29" .
```

Objects can be literals as in the age example, or other resources, as for the statement 'Alison has sibling Christopher' where 'Christopher' would be represented by another unique resource. Similarly, additional statements could be made about the 'Christopher' resource:

```
@prefix people: <http://example.org/people/> .
@prefix quality: <http://example.org/quality/> .
people:Alison quality:has_sibling people:Christopher .
people:Christopher quality:has_age "26" .
```

Much of human-generated data, particularly in the sciences, can be described using this basic triple structure. Linked Data is a paradigm for publishing data on the Web that uses RDF as a foundation for representation, and aims to make data a "first class citizen" of the Web to enable its widespread sharing, integration and re-use [12]. At the heart of Linked Data are these design principles: (1) use URIs to give globally scoped unique names to things, (2) use HTTP URIs so that those identifiers are resolvable on the Web, (3) describe things using Semantic Web standard languages so that the content at a resolved URI is machine understandable, and (4) include in those descriptions links to other things, so that the network of relationships between concepts can be traversed on the Web [12]. The query language of the Semantic Web is the SPARQL Protocol and RDF Query Language (SPARQL). RDF data is stored in a triplestore, a type of database

optimized for SPARQL querying and retrieval of triples. Triplestore implementations such as AllegroGraph, BigOWLIM, Sesame and Virtuoso enable access to distributed RDF resources on the Web at query time that can be integrated by federated queries.

For the life sciences in particular, Semantic Web technologies have emerged as a key enabling technology to tackle the challenge of information integration posed by the data-knowledge gap [13-15]. Specifically, RDF and Linked Data enable publishing experimental data on the Web in a manner that supports querying, automatic integration, and reasoning. The Bio2RDF project applies Linked Data principles to transform life sciences data from a variety of providers, and publishes nearly 30 billion triples of life science data through a globally redundant and distributed set of SPARQL endpoints [16, 17], using consistent resource-naming schemes to maximize integration across datasets. Recently developed data integration projects such as OpenPHACTS [18] consume Bio2RDF linked data and related bio-ontologies to facilitate discovery and reasoning.

1.5.1.2 Ontologies and the Web Ontology Language (OWL)

An ontology is the specification of a conceptualization [19] that describes concepts in a domain of knowledge and the relations between them. In the context of the Semantic Web (in contrast to philosophy, where ‘ontology’ has a related but distinct meaning), ontologies are serialized using a machine-understandable description language and are used to describe Linked Data entities and relationships between them. For example, an ontology for real estate might define concepts for ‘Property’, ‘House’, ‘Buyer’, ‘Seller’, ‘Sale price’, ‘Offer’ and so on, as well as relationships like ‘owns’ and ‘intends to purchase’, in such a way that a reasoner could automatically flag homes for sale that meet specific buyer criteria, without the requirement for a person to manually search all

available listings. An ontology for protein biochemistry would include concepts such as ‘Protein’, ‘Peptide’, ‘Amino acid’, ‘Beta sheet’, and relationships ‘has part’ and ‘encoded by’ and could enable the detection of secondary structures based on the existence of amino acids the have ontologically defined properties. The Web Ontology Language (OWL2) provides the means to describe classes (concepts), properties (relationships), individuals (instances of classes), and data values (such as raw numbers or strings). It also has elements such as existential and universal quantifiers to distinguish properties that may hold for some individuals of a class from those that must hold for all individuals of a class (*e.g. all* instances of the class ‘mother’ must have a child) , qualified cardinality restrictions to assert exactly how many entities or qualities an instance of class may be related to (*e.g.* a ‘mother’ has a *minimum* of 1 child), and class constructors (union, disjunction) to allow the description of complex classes as the union or disjunction of two or more other classes (*e.g.* a ‘mother’ is *equivalent to the union* of the classes ‘parent’ and ‘female’) [20, 21]. OWL2 is based on a family of Description Logics, a knowledge representation formalism that has its origins in semantic networks and frames [22]. OWL ontologies have been used to form the basis for developing reasoning-capable knowledge bases in the life sciences. As of March 2014, the National Center for Biomedical Ontology (NCBO) at Stanford University maintains a collection of over 380 bio-ontologies. One of the most widely used groups of ontologies is the Gene Ontology (GO) [23], which consists of three ontologies for biological processes, functions, and cellular locations. GO has been used to assign annotations to UniProt protein entries in the Gene Ontology Annotations (GOA) knowledge base [24], for gene set enrichment analysis [25], and numerous other applications including a recent effort to benchmark

protein function prediction approaches [26]. Biomedical ontologies have been used for integrating annotations from clinical notes to detect adverse drug events [27, 28]. SNOMED CT, a large medical ontology, has been used in a wide range of applications [29] including the classification of pathology reports to detect cancer diagnoses [30]. Other applications that take advantage of the reasoning capability of OWL ontologies includes work to describe genomic knowledge found in the *Saccharomyces* Genome Database (SGD) [31], the pharmacogenomics of depression as found in curated articles highlighted by the Pharmacogenomics Knowledge Base [32] and knowledge about RNA structure and function [33].

1.5.2 Hypothesis representation, formulation and evaluation

There have been many research efforts directed towards formulating and representing hypotheses and in computationally evaluating hypotheses using existing data. In this section I describe some of the exemplar systems and approaches in this field that are relevant to the life sciences, as well as their contributions and shortcomings.

HypGene [34, 35] was designed to describe and evaluate hypotheses about genetic attenuation. HypGene was implemented in Lisp and used theory revision operators to iteratively update hypotheses about the *trp* operon based on experimental data. Revision operators were triggered when predicted experimental outcomes generated by GENSIM [36] did not agree with observed experimental results. One example of a revision operator used by HypGene was an 'initial condition modification' operator, that updated a hypothesis by changing the amounts of input compounds present at the beginning of the hypothetical *in silico* experiment executed with GENSIM to satisfy the

observed outcomes. This necessarily required the expected outcomes of all relevant experiments to be specified in advance.

HinCyc [37] was used to predict the existence of biochemical pathways in *H. influenzae*. Specifically, HinCyc used data about *E. coli* pathways in the EcoCyc encyclopedia of *E. coli* genes and pathways [38] to hypothesize the occurrence of similar pathways in *H. influenzae*. To do this, HinCyc searched for homologs of known *E. coli* pathway components in the set of *H. influenzae* gene products, and hypothesized that a given pathway existed if a sufficient set of homologs was found.

GenePath [39], a system implemented in Prolog, used abductive reasoning to generate hypotheses about genetic networks based on genetic experiments in *Dictyostelium discoideum*. Abductive reasoning is a type of inference that proposes explanations (hypotheses) for a given observation. An example of abductive reasoning is to hypothesize that a gene's expression is increased based on the observation that there are higher levels of its protein product than observed in other conditions. In such cases, there may be multiple possible explanations for a given observation (consider an alternative hypothesis that a pathway which normally metabolizes the protein is not active), and which is the correct one is not necessarily known but also does not need to be for abductive reasoning to be useful in the context of biological hypothesis formulation. GenePath used background knowledge in the form of known genetic interactions (such as gene A induces expression of gene B, and gene C represses expression of gene D) and if-then rules to construct hypothetical genetic pathways to explain (satisfy the constraints of) a given set of experimental results.

Many years before this, Randolph Miller, Harry Pople Jr. and Jack Myer's work on the medical decision support system INTERNIST-I explored the potential of formal machine-understandable descriptions of medical knowledge for performing differential diagnoses [40]. INTERNIST-I was developed to encompass the knowledge held by medical doctor Jack Myer about internal medicine. It used a knowledge base that encompasses more than 15 person-years of curation work, and consisted of diseases, disease manifestations, and relationships between them including a disease hierarchy, as well as causality and predisposition relationships. INTERNIST-I operated by constructing differential diagnosis lists for a set of input patient conditions including laboratory results, physical features, and symptoms. Each differential diagnosis was then scored by taking into account several features of the diagnosis presentation, including the frequency with which a given symptom is observed with the diagnosis, how many symptoms are associated with a given possible diagnosis but are absent in the patient in question, how many presenting symptoms are not explained by the diagnosis. The diagnosis that scored significantly better than any other was proposed as the correct diagnosis. Despite the impressive effort involved in developing INTERNIST-I, and some positive evaluations, there have been significant critiques of this expert system and its shortcomings have been well described [41]. Its issues are related both to the structure and content of its knowledge base and also the logic used to reason over it for the purpose of differential diagnosis. For example, INTERNIST-I could not account for temporal relationships between manifestations, and could not distinguish between symptoms and causal factors of a diagnosed disease. It also generated false diagnoses that clinicians considered to be the result of basic reasoning error, and thus never saw widespread

application. Lastly, INTERNIST-I was not accessible to the novice user, even when the user was an expert clinician, because effective use required in-depth knowledge of its complex data input procedures as well as significant processing time.

HypGene, GenePath and INTERNIST-I are examples of rule-based systems applied to the problems of hypothesis generation and revision. Rule-based systems [42] consist of a collection of rules in a knowledge base (a 'rule base') and an interpreter or inference engine to execute rules triggered by input conditions. Rule-based systems have seen widespread use data and text mining, and in bioinformatics for a variety of applications including detecting [43] and modeling [44, 45] molecular pathways, extracting protein phosphorylation sites from the literature [46] and recognizing protein and gene names in scientific text [47]. Advantages of rule-based systems include that the modular nature of rules facilitates their reuse, new rules can be added to improve the scope and performance of rule bases, and that the ability to trace rule executions makes the reasoning of rule-based systems transparent to users [42].

More recently, Ross D. King and colleagues developed Adam the Robot Scientist [48], a combination system for carrying out automated wet lab experiments and reasoning over hypothesis spaces. Adam used abductive reasoning to formulate hypotheses about genes encoding 'orphan' enzymes (proteins which do not have a known corresponding gene) in yeast (*S. cerevisiae*), and deductive reasoning to test them by designing experiments and executing them automatically. Abductive reasoning was accomplished using a knowledge base consisting of a formal model of yeast metabolic pathways expressed in Prolog, and a database of known yeast genes and metabolites. King *et al.* used Adam to formulate and test 20 hypotheses about genes that may encode orphan

enzymes, and then performed additional experiments confirming three of the hypothesized gene-protein product relationships. An additional six hypotheses were confirmed by investigating the literature. The Robot Scientist uses a ontology to describe abduced hypotheses [49].

Tari *et al.* [50] developed a system that combines natural language processing of MedLine abstracts with a formal representation for drug-drug interactions (DDIs) in order to identify potentially undiscovered DDIs. Their system allowed for the formulation of hypothetical drug interactions and subsequent evaluation using drug interaction statements extracted from MedLine abstracts and DrugBank. Specifically, their system used a natural language processing (NLP) pipeline to extract parse trees that describe the syntactic (grammatical) structure of sentences and named entities (drugs, enzymes and genes) from MedLine abstracts, and stored these trees in a database. The database was then queried to extract explicit DDIs from processed sentences, as well as implicit DDIs inferred from statements (potentially occurring in different abstracts) such as 'drug1 inhibits enzymeA' and 'enzymeA metabolizes drug2'. Using AnsProlog, they executed rules concerning drug metabolism over the statements extracted from the abstracts to infer DDIs for a given drug. 20 of the 170 direct DDIs discovered by their method were reported in DrugBank, which was used as a gold standard.

RIBOWEB [51, 52] was an early Semantic Web tool developed in part by Russ Altman that represented scientific data about ribosomes in a formal machine understandable manner, and allowed users to evaluate by visual inspection three-dimensional models of ribosomes retrieved on the fly, using input from the user combined with reasoning over structural data. RIBOWEB used a knowledge base of

published structural and experimental data about the 30S ribosomal subunit, as well as the computational methods used to generate the structural data. The types of data and relationships between them are encoded in four ontologies about molecules, types of data, publications and methods.

The HyBrow (Hypothesis Browser) system [53, 54], developed by Stephen Racunas and Nigam Shah, was the inspiration and early guiding framework for HyQue. HyBrow used a manually curated knowledge base of literature-extracted facts about the galactose metabolism pathway in yeast (*Saccharomyces cerevisiae*) coupled with a model for hypotheses and rules to evaluate gene and protein-centric hypotheses about the genetic regulation of galactose metabolism in response to environmental cues. One of HyBrow's features is that it could rank proposed hypotheses by their degrees of support, and present this data to the user. HyBrow also had the ability to propose alternative hypotheses composed of events that were similar to hypothesized events but that had more support from the facts in the HyBrow knowledge base and violated fewer constraints. This capability was limited, however, to events that were already described in the knowledge base (*i.e.* alternative events were not composed on the fly by reasoning over their component parts). Another of its primary shortcomings, like INTERNIST-I, was the significant manual effort that was required to create the knowledge base at its core.

1.6 Summary

The approaches to hypothesis formulation and reasoning described above have made significant contributions in terms of methods for formally representing biological hypotheses and scientific data, but the implementation of these representation models is

typically system-specific and difficult to apply to new domains and integrate with other tools. Using Semantic Web standards and approaches for data integration to tackle these issues is a promising step forward. More importantly, the problem of *hypothesis formulation* is distinct from that of *hypothesis evaluation* and the majority of the reviewed literature aims to address the former. HyQue takes a fundamentally different approach from computational methods for hypothesis formulation, by framing the problem of hypothesis evaluation as one of data gathering and analysis in the context of domain knowledge. Hypothesis formulation tasks can be achieved only by executing pre-defined revision operators that can produce a limited set of possible alternatives and rely on *a priori* knowledge. By considering instead the problem of gathering and evaluating existing evidence for a given hypothesis, the HyQue framework is flexible and has the capability to evaluate hypotheses at any biological scale and using *any* kind of data, thereby taking advantage of the vast ocean of existing experimental data and expert knowledge in its fullest potential. It also addresses a need at the core of the biologist's work [55]: given a hypothesis a biologist already has, HyQue does the difficult work of retrieving and semi-automatically evaluating what we already know (but may not know to be *relevant*) in the context of a new idea. In the following chapters, I describe research I carried out to design and implement HyQue, and apply it to the task of evaluating outstanding biological hypotheses.

2 Chapter: Evaluating scientific hypotheses using the SPARQL Inferencing Notation (SPIN)

Abstract

Evaluating a hypothesis and its claims against experimental data is an essential scientific activity. However, this task is increasingly challenging given the ever growing volume of publications and data sets. Towards addressing this challenge, we previously developed HyQue, a system for hypothesis formulation and evaluation. HyQue uses domain-specific rulesets to evaluate hypotheses based on well understood scientific principles. However, because scientists may apply differing scientific premises when exploring a hypothesis, flexibility is required in both crafting and executing rulesets to evaluate hypotheses. Here, we report on an extension of HyQue that incorporates rules specified using the SPARQL Inferencing Notation (SPIN). Hypotheses, background knowledge, queries, results and now rulesets are represented and executed using Semantic Web technologies, enabling users to explicitly trace a hypothesis to its evaluation as Linked Data, including the data and rules used by HyQue. We demonstrate the use of HyQue to evaluate hypotheses concerning the yeast galactose gene system.

Contribution to thesis

In this chapter, I describe a major extension to the HyQue framework (first described in [56]) that satisfies Objective #1 of my thesis – designing and implementing a framework for hypothesis evaluation using Semantic Web standards and technologies. It also lays the groundwork for achieving Objective #2 by using the updated HyQue framework to evaluate biological hypotheses with well-documented experimental support, thereby acting as a proof of concept for the feasibility of a large-scale application of HyQue.

Permission to reproduce this published book chapter was granted by Springer Publishers: Callahan, A. & M. Dumontier. 2012. Evaluating scientific hypotheses using the SPARQL Inferencing Notation (SPIN). In Simperl, E. P. Cimiano, A. Polleres, O. Corcho & V. Presutti (Eds). The Semantic Web: Research and Applications, Lecture Notes in Computer Science Volume 7295, 2012, pp 647-658.

2.1 Introduction

Developing and evaluating hypotheses in the context of experimental research results is an essential activity for the life scientist, but one which is increasingly difficult to carry out manually given the ever growing volume of publications and data sets [2]. Indeed, biologists perceive that the predominant challenge in research is to "locate, integrate and access" the vast amounts of biological data resulting from small- and large-scale experiments [3]. Life sciences resources for the Semantic Web, such as Bio2RDF [16] and the growing number of bio-ontologies offer the potential to develop systems that consume these resources and computationally reason over the knowledge they contain to infer new facts [50, 57, 58] and answer complex questions [31].

With the diversity of research claims that exist in such large resources, there is also the potential for statements to contradict one another. Formally exploring the outcomes of relying on different sets of research claims to assess a hypothesis is necessary to not only confer confidence in the hypothesis evaluation methodology (whether manual or automatic), but also to provide evidence for the likelihood of one interpretation of results compared to another. Previous research efforts that have aimed at formally evaluating scientific data in the context of hypotheses include HypGene [34, 35], HinCyc [37], GenePath [39] and Adam the Robot Scientist [48, 49]. Each of these projects use application-specific representations for data and the rules used to assess this data, making their extension to new domains, as well as their comparison and performance evaluation difficult.

Towards addressing the challenge of integrating experimental knowledge with biological hypotheses, we previously developed HyQue [59, 60]. HyQue uses Semantic

Web standard languages (RDF/OWL) to represent hypotheses and data, SPARQL queries to retrieve data, and domain-specific rulesets to evaluate hypotheses against this data. While HyQue uses rulesets based on well understood scientific principles [53, 61], finer grained evaluations would require the exclusion or inclusion of additional rules. Problematically, HyQue's domain-specific evaluation rules were hard-coded, which made it implausible for users to construct custom rule sets for hypothesis evaluation.

In this paper, we describe an extension of HyQue that uses evaluation rules specified using the SPARQL Inferencing Notation (SPIN) in place of hardcoded rules. SPIN is a W3C member submission[1] rule language whose scope and expressivity are defined by SPARQL. Thus, SPIN rules are SPARQL queries which can not only be used to assert new facts, but also used to infer OWL class membership for non-hierarchical class membership axioms[2]. Moreover, SPIN rules can be serialized into RDF, and hence can become part of a system that maintains provenance concerning calculations and inferences.

In this new version of HyQue, hypotheses, background knowledge, queries, results and now evaluation rulesets are represented and executed using Semantic Web technologies. Domain specific rules for evaluating experimental data in the context of a hypothesis are now maintained independently of the system rules that are used to calculate overall hypothesis evaluation scores. We demonstrate these features by evaluating hypotheses about the galactose gene system in yeast [61]. HyQue enables users to explicitly trace a hypothesis to its evaluation, including the data and rules used. In addition to making the hypothesis evaluation methodology transparent and

---

[1] http://www.w3.org/Submission/2011/SUBM-spin-overview-20110222/
[2] http://www.w3.org/Submission/2011/SUBM-spin-modeling-20110222/

reproducible (essential qualities for good e-science), this allows scientists to discover experimental data that support a given hypothesis as well as explore new and potentially uncharacterized links between multiple research outcomes. A unique strength of HyQue is that its design is not dependent upon a specific biological domain, and the assumptions encoded in its hypothesis evaluation rules are changeable and maintained separately from the evaluation system. As our understanding of biological systems evolves and improves through research, the way HyQue evaluates hypotheses, as well as the facts and data it uses, can evolve as well.

2.2 Methods

2.2.1 Overview

HyQue evaluates hypotheses (and assigns an evaluation score) by executing SPIN rules over the pertinent knowledge extracted from a HyQue Knowledge Base (HKB). A hypothesis is formulated as a logical expression in which elements of the hypothesis correspond to biological entities of interest. HyQue maps the hypothesis, expressed using terminology from the HyQue ontology[3], to the relevant SPIN rules, which execute SPARQL queries to retrieve data from the HKB. Finally, HyQue executes additional SPIN rules over the extracted data to obtain a quantitative measure of hypothesis support. Figure 3 provides a graphical overview of HyQue.

[3] The HyQue ontology, linked data, and SPIN rules are available at the project website: http://hyque.semanticscience.org

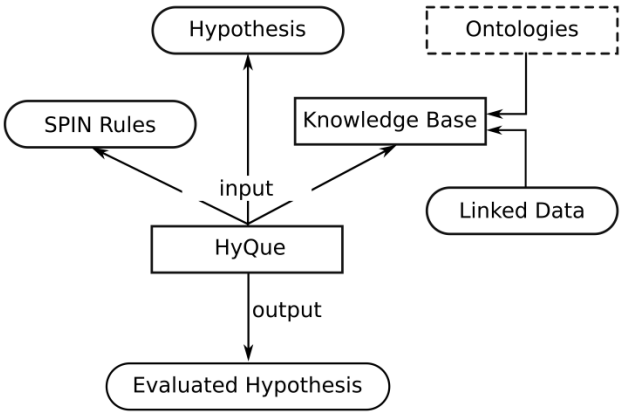


Figure 3 HyQue uses SPIN rules to evaluate a hypothesis over RDF linked data and OWL ontologies. The dashed rectangle represents OWL ontologies. Rounded rectangles are RDF resources.

2.2.2 HyQue hypothesis model

A HyQue *hypothesis* may be composed of one or more *propositions* that *specify events* related to each other by AND/OR operators. Events must have an agent (an entity executing an action) and a target (the object of the action), and can optionally have a physical location, a physical operator (*e.g.* 'binding'), a logical operator (*e.g.* 'repression' or 'activation') and a perturbation context (in the case of genes and proteins). HyQue maps these events to SPARQL queries through a SPIN rule, and subsequently executes them over the HyQue Knowledge Base. HyQue currently supports the following kinds of events [59, 60]:

1. protein-protein binding
2. protein-nucleic acid binding
3. molecular activation
4. molecular inhibition
5. gene induction
6. gene repression
7. transport

2.2.3 HyQue Knowledge Base (HKB)

A HyQue Knowledge Base (HKB) consists of RDF data, RDFS-based class hierarchies and/or OWL ontologies. For demonstration purposes, our HKB consists of an RDF version of the galactose (GAL) gene network in yeast [53], an extended version of the Bio2RDF compatible yOWL knowledge base [31, 56] and the following bio-ontologies (for the listed entities):

- **Gene Ontology (GO)**: cellular components, events (*e.g.* 'nucleus', 'positive regulation of gene expression')
- **Evidence Codes Ontology (ECO)**: the type of evidence supporting an event (*e.g.* 'electronic annotation', 'direct assay')
- **Sequence Ontology (SO)**: event participants (*e.g.* 'gene')
- **Chemical Entities of Biological Interest (CHEBI) Ontology**: event participants (*e.g.* 'protein', 'galactose')

All Linked Data (encoded using RDF) and ontologies (encoded using OWL) that comprise the HKB are available at the project website.

2.2.4 The HyQue scoring system

HyQue uses rules to calculate a numerical score for a hypothesis based on the degree of support the hypothesis has from statements in the HKB. HyQue first attempts to identify statements about experimentally verified events in the HKB that have a high degree of matching to a hypothesized event, and then assesses these statements using domain specific rules to assign a score to the hypothesized event. If there is a statement about an experimentally reported GAL gene/protein interaction in the HKB that exactly matches a hypothesized event, then that event will be assigned a maximum score when it is

evaluated by HyQue. In contrast, if a hypothesized event describes an interaction between a protein A and a protein B but there is a statement in the HKB asserting that protein A does *not* interact with protein B, then the hypothesis will be assigned a low score based on the negation of the hypothesized event by experimental data. Different HyQue rules add or subtract different numerical values based on whether the relevant experimental data has properties that provide support for a hypothesized event. For instance, if an event is hypothesized to occur in a specific cellular compartment *e.g.* nucleus, but the HKB only contains a statement that such an event takes place in a different cellular component *e.g.* cytoplasm, then a rule could be formulated such that the hypothesis, while not directly supported by experimental evidence, will be penalized less than if the event had been asserted to not take place at all.

Based on such scoring rules, each event type has a maximum possible score. When a hypothesized event is evaluated by HyQue, it is assigned a normalized score calculated by the sum of the output of the relevant rule(s) divided by the maximum possible score. In this way, if an event has full experimental support, it will have an overall score of 1, while if only some properties of the hypothesized event are supported by statements in the HKB is will have a score between 0 and 1.

Overall proposition and hypothesis scores are calculated by additional rules based on the operators that relate events. If a proposition specifies ‘event A’ OR ‘event B’ OR ‘event C’ then the maximum event score will be assigned as the proposition score, while if the ‘AND’ operator was used, the mean event score will be assigned as the proposition score. Using the mean reflects the relative contribution of each event score while still

maintaining a normalized value between 0 and 1. Similar rules are used to calculate an overall hypothesis score based on proposition scores.

HyQue uses SPIN to execute rules that reflect this scoring system.

2.2.5 HyQue SPIN rules

HyQue uses two types of rules to evaluate hypotheses: *domain specific rules* that depend on the subject of the hypothesis (in this case, gene regulation) and *system rules* that define how to combine the output of domain specific rules in order to determine an overall hypothesis evaluation score. These rules are defined separately using SPIN and can be changed independently of each other.

HyQue system rules describe how to calculate event, proposition and overall hypothesis scores based on the structure and content of the hypothesis. For example, the following rule (modified with single quoted labels for illustrative purposes) generates four statements that assert the relationship between a HyQue hypothesis (any instance of the class hyque:HYPOTHESIS_0000000) and its evaluation.

```
CONSTRUCT {
   ?this ‘has attribute’ ?hypothesisEval .
   ?hypothesisEval a ‘evaluation’.
   ?hypothesisEval ‘obtained from’ ?propositionEval .
   ?hypothesisEval ‘has value’ ?hypothesisEvalScore .
} WHERE {
   ?this ‘has component part’ ?proposition .
   ?proposition ‘has attribute’ ?propositionEval .
   BIND(:calculateHypothesisScore(?this) AS ?hypothesisEvalScore) .
      BIND(IRI(fn:concat(afn:namespace(?this), afn:localname(?this),"_", "evaluation"))
      AS ?hypothesisEval) .
}
```

This SPIN rule states that a HyQue hypothesis (hyque:HYPOTHESIS_0000000) will be related to a new attribute of type ‘evaluation’ (hyque:HYPOTHESIS_0000005) by the

'has attribute' (hyque:HYPOTHESIS_0000008) object property. The numeric value of this evaluation is specified using the 'has value' (hyque:HYPOTHESIS_0000013) datatype property. Since the evaluation of the hypothesis comes from evaluating the propositional parts, these are related with the 'is obtained from' (hyque:HYPOTHESIS_0000007) object property. The SPARQL variable '?this' has a special meaning for SPIN rules, and refers to any instance of the class the rule is linked to. SPIN rules are linked to classes in the HyQue ontology using the spin:rule predicate.

This hypothesis rule uses another rule, calculateHypothesisScore, to calculate the hypothesis score, and the output of executing this rule is bound to the variable ?hypothesisEvalScore. Note that the hypothesis rule is constrained to a HyQue hypothesis that 'has component part' (hyque:HYPOTHESIS_0000010) some 'proposition' (hyque:HYPOTHESIS_0000001) that 'has attribute' a proposition evaluation. In this way HyQue rules are chained together – when one rule is executed, all the rules it depends on are executed until no new statements are created. In this case, because a hypothesis evaluation score requires a proposition evaluation score, when the hypothesis evaluation rule is executed, the HyQue SPIN rule for calculating a proposition score is executed as well. Each proposition evaluation is asserted to be 'obtained from' the event evaluations corresponding to the event(s) specified by (hyque:HYPOTHESIS_0000012) the proposition. Each event evaluation is also asserted to be 'obtained from' the scores determined for each event property (the agent, target, location *etc.*) and the statements in the HKB the scores are based on.

Domain specific rules for HyQue pertain to the domain of interest. An example of a domain specific rule is calculateActivateEventScore corresponding to the following SPARQL query:

```
SELECT ?activateEventScore
WHERE {
        BIND (:calculateActivateAgentTypeScore(?arg1)
                AS ?agentTypeScore) .
        BIND (:calculateActivateTargetTypeScore(?arg1)
                AS ?targetTypeScore) .
        BIND (:calculateActivateLogicalOperatorScore(?arg1)
                AS ?logicalOperatorScore) .
        BIND (:penalizeNegation(?arg1) AS ?negationScore) .
        BIND (3 AS ?maxScore) .
        BIND (((((?agentTypeScore + ?targetTypeScore) +
                        ?logicalOperatorScore) + ?negationScore) /
                        ?maxScore) AS ?activateEventScore) .
}
```

In this rule, a numeric score (?activateEventScore) is calculated from the sum of a set of outputs from other sub-rules divided by the maximum score possible (in this case, 3). This rule uses a special variable ?arg1, which corresponds to any entities linked using the SPIN sp:arg1 predicate. This special variable is selected by specifying a spin:constraint on the rule, which states that any variable passed to the rule when it is called can be referred to within the rule to by '?arg1'. For example, if the rule were called by including calculateActivateEventScore(?data) in a SPARQL query WHERE statement, ?data will be the variable referenced by ?arg1 in the rule definition.

The sub-rule calculateActivateLogicalOperatorScore determines a score for the type of logical operator specified in a HyQue hypothesis based on domain specific knowledge about the GAL gene network. This rule corresponds to the following SPARQL query:

```
SELECT ?score
WHERE {
?arg1 'has logical operator' ?logical_operator .
BIND (IF((?logical_operator = 'positive regulation of molecular
        function'), 1, -1) AS ?score) .
}
```

Thus, if the logical operator specified in a hypothesis event is of type 'positive regulation of molecular function' (GO:0044093) the rule will return 1, and otherwise the rule will return -1.The calculateActivateEventScore rule is composed of several sub-rules of this format. HyQue uses similar rules for each of the seven event types listed in section 2.2.2 to evaluate hypotheses.

SPIN rules were composed using the free edition of TopBraid Composer 3.5. HyQue executes SPIN rules using the open source SPIN API 1.2.0 and Jena 2.6.4.

2.2.6 Executing HyQue SPIN rules over the HKB

To execute the HyQue SPIN rules over an input hypothesis using data from the HKB, a Java program was written with the open source SPIN API (version 1.2.0) and the Jena API (version 2.6.4). Users can submit a hypothesis to the program via a servlet available at http://hyque.semanticscience.org. The servlet returns the RDF-based hypothesis evaluation.

2.3 Results

HyQue currently uses a total of 63 SPIN rules to evaluate hypotheses. 18 of these are system rules, and the remaining 45 are domain specific rules that calculate evaluation scores based on well understood principles of the GAL gene network in yeast as described in section 2.2.5. These rules have been used to evaluate 5 representative hypotheses about the GAL domain, one of which is presented in detail in section 2.3.1.

2.3.1 Evaluating a hypothesis about GAL gene induction and protein inhibition

The following is a natural language description of a hypothesis about the GAL gene network that has been evaluated by HyQue. Individual events are indicated by the letter 'e', followed by a number to uniquely identify them. Events are related by the AND operator in this hypothesis, while the two sets of events (typed as propositions in the HyQue hypothesis ontology) are related by the OR operator.

(Gal4p induces the expression of GAL1 *e1*

AND Gal3p induces the expression of GAL2 *e2*

AND Gal4p induces the expression of GAL7) *e3*

OR

(Gal4p induces the expression of GAL7 *e4*

AND Gal80p induces the expression of GAL7 *e5*

AND Gal80p does not inhibit the activity of Gal4p

when GAL3 is over-expressed) *e6*

Two domain specific SPIN rules were executed to evaluate this hypothesis: calculateInduceEventScore for *e1-e5* and calculateInhibitEventScore for *e6,* in conjunction with system rules to calculate overall proposition and hypothesis scores based on the event scores.

By identifying and evaluating statements in the HKB that experimentally support *e1*, the calculateInduceEventScore rule assigns *e1* a score of 4 out of a maximum score of 5 (see Table 1). This corresponds to a normalized score of 0.8. Similarly, events 2-5 also receive a score of 0.8. The calculateInhibitEventScore rule assigns event 6 a score of 1 based on comparable scoring rules. Therefore, the proposition specifying *e4, e5* and

e6 receives a higher score (0.87 – the mean of the individual event scores) than the proposition specifying *e1, e2* and *e3* (with a mean score of 0.8). Because the two propositions were related by the OR operator, the hypothesis is assigned an overall score that is the maximum of the two proposition scores, in this case, a value of 0.87.

Table 1 SPIN rules executed to evaluate a hypothetical GAL gene induction event, their outcomes, and contribution to an overall hypothesis score assigned by HyQue

SPIN Rule	Rule output	Score
penalizeNegation	Event is not negated	0
calculateInduceAgentTypeScore	Actor is a 'protein' (CHEBI:36080)	+1
calculateInduceTargetTypeScore	Target is a 'gene' (SO:0000236)	+1
calculateInduceLogical OperatorScore	Logical operator is 'induce' (GO:0010628)	+1
calculateInduceAgentFunction Score	Actor does not have 'transcription factor activity' (GO:0003700)	0
calculateInduceLocationScore	Location is 'nucleus' (GO:0005634)	+1

The complete HyQue evaluations of this hypothesis as well as that of four additional hypotheses are available as RDFat the project website.

2.3.2 Changing a domain specific rule affects hypothesis evaluation

The calculateInhibitEventScore used to evaluate event 6 in section 2.3.1 in its current form does not take into account the physical location of the event participants. In other words, the score does not depend on data describing where the event participants are known (or not) to be located in the cell. However, some experimental evidence suggests that physical location in the context of an inhibition event plays an important role. Specifically, the inhibition of Gal4p activity by Gal80p is known to take place in the nucleus, yet this inhibition is interrupted when Gal80p is bound by Gal3p, which is typically found in the cytoplasm [62].

The effect of changing the calculateInhibitEventScore rule to require that all event participants be located in the nucleus to achieve a maximum score (a reasonable assumption given published findings [63]) on the hypothesis in section 2.3.1 would be that the score for *e6* is reduced. This is because adding an additional sub-rule (let us call it calculateInhibitEventParticipantLocationScore) would increase the maximum score, while experimental data in the HKB is not available to satisfy the conditions of this new sub-rule – there is not experimental data available about the location of the Gal4p or Gal80p proteins in the cell. More specifically, let us say that the maximum score possible for calculateInhibitEventScore with the new sub-rule is now 4, and that event 6 is therefore assigned a score of 0.75 (3/4) based on the output of this rule. This changes the overall hypothesis score in that the first proposition (specifying events 1-3) now has a higher mean score (0.8, *versus* 0.78 for the second proposition as calculated using the new rule), and thus this is assigned as the overall hypothesis score.

This example demonstrates how using a different domain specific rule affects an overall hypothesis evaluation, and how the effect can be traced to both the rule(s) used and the data the rules are executed over.

2.4 Discussion

Using SPIN rules to evaluate HyQue hypotheses has several advantages. While HyQue "version 1.0" used SPARQL queries to obtain relevant statements from the HKB, the scoring rules used to evaluate those statements were hard-coded in system code. HyQue's SPIN evaluation rules can be represented as RDF, which allows the potential for users to query for HyQue rules that meet specific conditions, as well as potentially link to and aggregate those rules. In addition, users can create their own SPIN rules to meet specific

evaluation criteria and augment existing HyQue rules to include them. In this way, different scientists may use the same data to evaluate the same hypotheses and arrive at unique evaluations depending on the domain principles encoded by the SPIN rules they use, as demonstrated in section 2.3.2. Encoding evaluation criteria as SPIN rules also ensures that the source of an evaluation can be explicitly stated, both in terms of the rules executed and the data the rules were executed over. This is crucial for formalizing the outcomes of scientific reasoning such that research conclusions can be confidently stated.

Separating HyQue system rules from the GAL domain specific rules highlights the two aspects of the HyQue scoring system. Specifically, HyQue currently encodes certain assumptions about how events in hypotheses may be related to one another, and how these relations are used to determine an overall hypothesis score, as well as domain specific assumptions about how to evaluate data in the context of knowledge about the GAL gene network. However, because assumptions about hypothesis structure are encapsulated by HyQue system rules, they may be changed or augmented without affecting the GAL domain specific rules, and *vice versa*. HyQue system rules can be extended over time to facilitate the evaluation of hypotheses that have fundamentally different structures than those currently presented as demonstrations. We envision a future iteration of HyQue where users can submit unique system and domain specific rules to use for evaluating hypotheses and in this way further research in their field by exploring novel interpretations of experimental data and hypotheses. Similarly, it may be possible in future for HyQue users to select from multiple sets of evaluation rules and to compare the hypothesis evaluations that result.

Crafting SPIN rules requires knowledge of SPARQL, which, while being used in a number of life-science related projects [14, 16, 57, 64, 65], may present a barrier to some users. Similarly, representing hypotheses as RDF to submit to HyQue is not a trivial activity. To address the latter, we have developed an online form based system for specifying hypothesis details and converting them to RDF, available at the project website.

The Rule Interchange Format (RIF)[4] is the W3C standard for representing and exchanging rules between rule systems. SPIN, a W3C member submission, has been identified as an effort complementary to RIF[66] and because there is some discussion of RIF and RDF compatibility[5], SPIN and RIF may become compatible if the RIF working group remains active[6]. HyQue provides a relevant use case and motivation for enabling such compatibility. Given that SPIN rules may be represented as RDF and executed over any RDF store using SPARQL (both W3C standards), however, and that the motivation of SPIN is specifically to execute SPARQL as rules, in the context of HyQue compatibility with RIF is not of immediate concern.

2.5 Conclusions

We present an extended version of HyQue that uses SPIN rules to evaluate hypotheses encoded as RDF, and makes the evaluation, including the data it is based upon, also available as RDF. In this way, users are able to explicitly trace a path from hypothesis to evaluation and the supporting experimental data, and *vice versa*. We have demonstrated how HyQue evaluates a specific hypothesis about the GAL gene network in yeast with an

---

[4] http://www.w3.org/TR/2010/NOTE-rif-overview-20100622/
[5] http://www.w3.org/TR/2010/REC-rif-rdf-owl-20100622/
[6] http://www.w3.org/Submission/2011/02/Comment/

explanation of the scoring rules used and their outcomes. Evaluations of additional hypotheses, as well as HKB data and HyQue SPIN rules are available at http://hyque.semanticscience.org.

3 Chapter: Bio2RDF Release 2 - Improved coverage, interoperability and provenance of Life Science Linked Data

Abstract

Bio2RDF currently provides the largest network of Linked Data for the Life Sciences. Here, we describe a significant update to increase the overall quality of RDFized datasets generated from open scripts powered by an API to generate registry-validated IRIs, dataset provenance and metrics, SPARQL endpoints, downloadable RDF and database files. We demonstrate federated SPARQL queries within and across the Bio2RDF network, including semantic integration using the Semanticscience Integrated Ontology (SIO). This work forms a strong foundation for increased coverage and continuous integration of data in the life sciences.

Contribution to thesis

A core requirement for HyQue is having access to large amounts of structured biological data for hypothesis evaluation, and thus the work described in this Chapter of improving the Bio2RDF approach to generating biological linked data and extending the Bio2RDF network with new datasets was a key aspect of my doctoral research. Specifically, the linked data made available via Bio2RDF Release 2 (>3 billion statements as of March 2014) for querying was essential for successfully applying HyQue as described in Chapters 2 and 5, towards achieving Objective #2. Secondly, in Chapter 5 I describe the addition of three new datasets to the Bio2RDF linked data network for the purpose of applying HyQue to the domain of aging in *C. elegans*, and this Chapter describes the methodology used for generating and publishing Bio2RDF linked data in detail.

Permission to reproduce this published book chapter was granted by Springer Publishers:

Callahan, A, J. Cruz-Toledo, P. Ansell & M. Dumontier. 2013. Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In Cimiano, P., O. Corcho, V. Presutti, L. Hollink, & S. Rudolph (Eds). *The Semantic Web: Semantics and Big Data,* Lecture Notes in Computer Science Volume 7882, 2013, pp 200-212.

3.1 Introduction

With the advent of the World Wide Web, journals have increasingly augmented their peer-reviewed journal publications with downloadable experimental data. While the increase in data availability should be cause for celebration, the potential for biomedical discovery across all of these data is hampered by access restrictions, incompatible formats, lack of semantic annotation and poor connectivity between datasets [3]. Although organizations such as the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) have made great strides to extract, capture and integrate data, the lack of formal, machine-understandable semantics results in ambiguity in the data and the relationships between them. With over 1500 biological databases, it becomes necessary to implement a more sophisticated scheme to unify the representation of diverse biomedical data so that it becomes easier to integrate and explore [67]. Importantly, there is a fundamental need to capture the provenance of these data in a manner that will support experimental design and reproducibility in scientific research. Providing data also presents real practical challenges, including ensuring persistence, availability, scalability, and providing the right tools to facilitate data exploration including query formulation.

The Resource Description Framework (RDF) provides an excellent foundation to build a unified network of linked data on the emerging Semantic Web. While an increasing number of approaches are being proposed to describe and integrate specific biological data [68-70], it is the lack of coordinated identification, vocabulary overlap and alternative formalizations that challenges the promise of large-scale integration [71]. Formalization of data into ontologies using the Web Ontology Language (OWL) have

yielded interesting results for integration, classification, consistency checking and more effective query answering with automated reasoning [62, 72-75]. However, these efforts build the ontology in support of the task and there is little guarantee that the formalization will accommodate future data or support new applications. Alternatively, integration of data may be best facilitated by independent publication of datasets and their descriptions and subsequent coordination into integrative ontologies or community standards. This approach provides maximum flexibility for publishing original datasets with publisher provided descriptors in that they are not constrained by limited standards, but provides a clear avenue for future integration into a number of alternative standards.

Bio2RDF is a well-recognized open-source project that provides linked data for the life sciences using Semantic Web technologies. Bio2RDF scripts convert heterogeneously formatted data (*e.g.* flat-files, tab-delimited files, dataset specific formats, SQL, XML etc.) into a common format – RDF. Bio2RDF follows a set of basic conventions to generate and provide Linked Data which are guided by Tim Berners-Lee's design principles[7] , the Banff Manifesto[8] and the collective experience of the Bio2RDF community. Entities, their attributes and relationships are named using a simple convention to produce Internationalized Resource Identifiers (IRIs) while statements are articulated using the lightweight semantics of RDF Schema (RDFS) and Dublin Core. Bio2RDF IRIs are resolved through the Bio2RDF Web Application, a servlet that answers Bio2RDF HTTP requests by formulating SPARQL queries against the appropriate SPARQL endpoints.

---

[7] http://www.w3.org/DesignIssues/Principles.html
[8] https://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto

Although several efforts for provisioning linked life data exist such as Neurocommons [76], LinkedLifeData [77], W3C HCLS[9] , Chem2Bio2RDF [78] and BioLOD, Bio2RDF stands out for several reasons: i) Bio2RDF is open source and freely available to use, modify or redistribute, ii) it acts on a set of basic guidelines to produce syntactically interoperable linked data across all datasets, iii) does not attempt to marshal data into a single global schema, iv) provides a federated network of SPARQL endpoints and v) provisions the community with an expandable global network of mirrors that host RDF datasets. Thus, Bio2RDF uniquely offers a community-focused resource for creating and enhancing the quality of biomedical data on the Semantic Web.

Here, we report on a second coordinated release of Bio2RDF, Release 2 (R2), which yields substantial increases in syntactic and semantic interoperability across refactored Bio2RDF datasets. We address the problem of IRI inconsistency arising from independently generated scripts through an API over a dataset registry to generate validated IRIs. We further generate provenance and statistics for each dataset, and provide public SPARQL endpoints, downloadable database files and RDF files. We demonstrate federated SPARQL queries within and across the Bio2RDF network, including queries that make use of the Semanticscience Integrated Ontology (SIO)[10] , which provides a simple model with a rich set of relations to coordinate ontologies, data and services.

3.2 Methods

[9] http://www.w3.org/blog/hcls/
[10] http://code.google.com/p/semanticscience/wiki/SIO

In the following section we will discuss the procedures and improvements used to generate Bio2RDF R2 compliant Linked Open Data including entity naming, dataset provenance and statistics, ontology mapping, query and exploration.

3.2.1 Entity naming

For data with a source assigned identifier, entities are named as follows:

http://bio2rdf.org/namespace:identifier

where 'namespace' is the preferred short name of a biological dataset as found in our dataset registry and the 'identifier' is the unique string used by the source provider to identify any given record. For example, the HUGO Gene Nomenclature Committee identifies the human prostaglandin E synthase gene (PIG12) with the accession number "9599". This dataset is assigned the namespace "hgnc" in our dataset registry, thus, the corresponding Bio2RDF IRI is

http://bio2rdf.org/hgnc:9599

For data lacking a source assigned identifier, entities are named as follows:

http://bio2rdf.org/namespace_resource:identifier

where 'namespace' is the preferred short name of a biological dataset as found in our dataset registry and 'identifier' is uniquely created and assigned by the Bio2RDF script. This pattern is often used to identify objects that arise from the conversion of n-ary relations into an object with a set of binary relations. For example, the Comparative Toxicogenomics Database (CTD) describes associations between diseases and drugs, but does not specify identifiers for these associations, and hence we assign a new stable identifier for each, such as

http://bio2rdf.org/ctd_resource:C112297D029597

for the chemical-disease association between 10,10-bis(4-pyridinylmethyl)-9(10H)-anthracenone (mesh:C112297) and the Romano-Ward Syndrome (mesh:D029597).

Finally, dataset-specific types and relations are named as follows:

http://bio2rdf.org/namespace_vocabulary:identifier

where 'namespace' is the preferred short name of a biological dataset as found in our dataset registry and 'identifier' is uniquely created and/or assigned by the Bio2RDF script. For example, the NCBI's HomoloGene resource provides groups of homologous eukaryotic genes and includes references to the taxa from which the genes were isolated. Hence, the Homologene group is identified as a class

http://bio2rdf.org/homologene_vocabulary:HomoloGene_Group

while the taxonomic relation is specified with

http://bio2rdf.org/homologene_vocabulary:has_taxid

3.2.2 Open source scripts

In 2012, we consolidated the set Bio2RDF open source[11] scripts into a single GitHub repository (bio2rdf-scripts[12]). GitHub facilitates collaborative development through project forking, pull requests, code commenting, and merging. Thirty PHP scripts, one Java program and a Ruby gem are now available for any use (including commercial), modification and redistribution by anyone wishing to generate Bio2RDF data, or to improve the quality of RDF conversions currently used in Bio2RDF.

3.2.3 Programmatically accessible resource registry

[11] http://opensource.org/licenses/MIT
[12] https://github.com/bio2rdf/bio2rdf-scripts

In order to ensure consistency in IRI assignment by different scripts, we established a common resource registry that each script must make use of. The resource registry specifies a unique namespace for each of the datasets (a.k.a. namespace; e.g. 'pdb' for the Protein Data Bank), along with synonyms (e.g. ncbigene, entrez gene, entrezgene/locuslink for the NCBI's Gene database), as well as primary and secondary IRIs used within the datasets (e.g. http://purl.obolibrary.org/obo/, http://purl.org/obo/owl/, http://purl.obofoundry.org/namespace, etc.) when applicable. The use of the registry in this way ensures a high level of syntactic interoperability between the generated linked data sets.

3.2.4 Provenance

Bio2RDF scripts now generate provenance using the Vocabulary of Interlinked Datasets (VoID), the Provenance vocabulary (PROV) and Dublin Core vocabulary. As illustrated in Figure 4, each item in a dataset is linked using void:inDataset to a provenance object (typed as void:Dataset). The provenance object represents a Bio2RDF dataset, in that it is a version of the source data whose attributes include a label, the creation date, the creator (script URL), the publisher (Bio2RDF.org), the Bio2RDF license and rights, the download location for the dataset and the SPARQL endpoint in which the resource can be found. Importantly, we use the W3C PROV relation 'wasDerivedFrom' to link this Bio2RDF dataset to the source dataset, along with its licensing and source location.

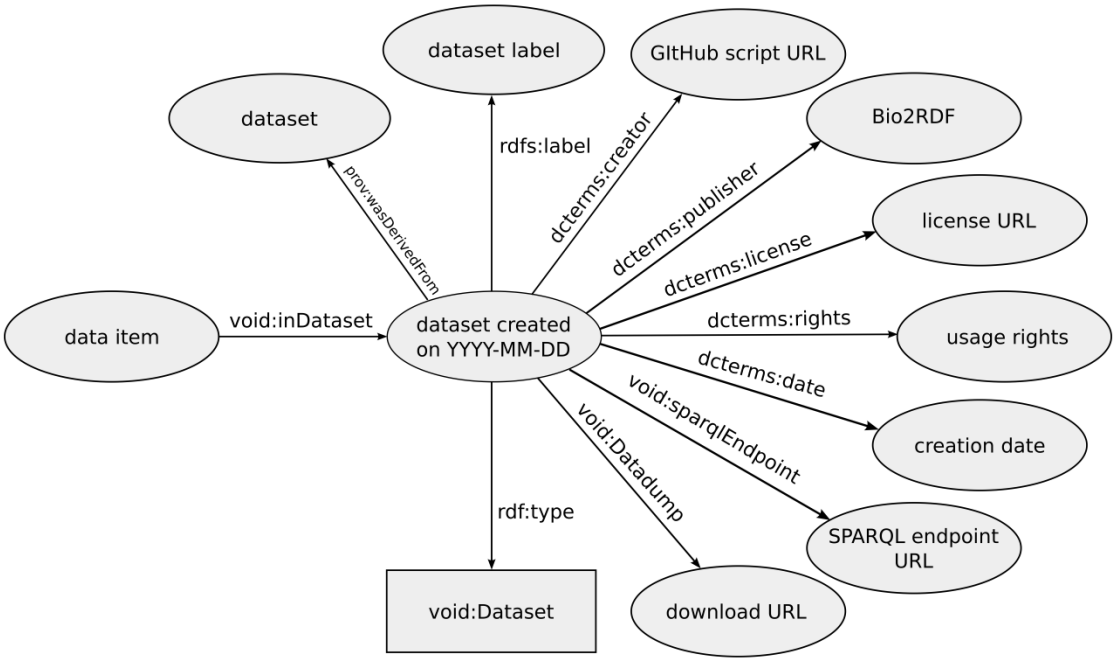


Figure 4 The Bio2RDF R2 provenance model.

3.2.5 Dataset metrics

A set of nine dataset metrics are computed for each dataset that summarize their contents:

- total number of triples
- number of unique subjects
- number of unique predicates
- number of unique objects
- number of unique types
- number of unique objects linked from each predicate
- number of unique literals linked from each predicate
- number of unique subjects and objects linked by each predicate
- unique subject type-predicate-object type links and their frequencies

These metrics are serialized as RDF using our own vocabulary using the namespace

http://bio2rdf.org/dataset_vocabulary, and subsequently loaded into a named graph at each dataset SPARQL endpoint with the following pattern:

http://bio2rdf.org/bio2rdf-*namespace*-statistics

where namespace is the preferred short name for the Bio2RDF dataset. While the values for metrics 1-4 are provided via suitably named datatype properties, metric s 5-9 require a more complex, typed object. For instance, a SPARQL query to retrieve all type-predicate-type links and their frequencies from the CTD endpoint is:

```
PREFIX statistics: <http://bio2rdf.org/dataset_vocabulary:>
SELECT *
FROM <http://bio2rdf .org/bio2rdf-ctd-statistics>
WHERE {
        ?endpoint a statistics:Endpoint.
        ?endpoint statistics:has_type_relation_type_count ?c.
        ?c statistics:has_subject_type ?subjectType.
        ?c statistics:has_subject_count ?subjectCount.
        ?c statistics:has_predicate ?predicate.
        ?c statistics:has_object_type ?objectType.
        ?c statistics:has_object_count ?objectCount.
}
```

Furthermore, to support context-sensitive SPARQL query formulation using SparQLed [79], we generated the data graph summaries using the Dataset Analytics Vocabulary[13]. These are stored in each endpoint in the graph named http://sindice.com/analytics.

3.2.6 Bio2RDF to SIO ontology mapping

Since each Bio2RDF dataset is expressed in terms of a dataset-specific vocabulary for its types and relations, it becomes rather challenging to compose federated queries across both linked datasets as well as datasets that overlap in their content. To facilitate dataset-independent querying, Bio2RDF dataset-specific vocabulary were mapped to the

[13] http://vocab.sindice.net/analytics#

Semanticscience Integrated Ontology (SIO), which is also being used to map vocabularies used to describe SADI-based semantic web services. Dataset specific types and relations were extracted using SPARQL queries and manually mapped to corresponding SIO classes, object properties and datatype properties using the appropriate subclass relation (*i.e.* rdfs:subClassOf, owl:SubObjectPropertyOf). Bio2RDF dataset vocabularies and their SIO-mappings are stored in separate OWL ontologies on the bio2rdf-mapping GitHub repository[14].

3.2.7 SPARQL endpoints

Each dataset was loaded into a separate instance of OpenLink Virtuoso Community Edition version 6.1.6 with the faceted browser, SPARQL 1.1 query federation and Cross-Origin Resource Sharing (CORS) enabled.

3.2.8 Bio2RDF web application

Bio2RDF Linked Data IRIs are made resolvable through the Bio2RDF Web Application, a servlet based application that uses the QueryAll Linked Data library [80] to dynamically answer requests for Bio2RDF IRIs by aggregating the results of SPARQL queries to Bio2RDF SPARQL endpoints that are automatically selected based on the structure of the query IRI. The Web Application can be configured to resolve queries using multiple SPARQL endpoints, each of which may handle different namespaces and identifier patterns. Such configurations are stored as RDF, and specified using Web Application profiles. Profiles are designed to allow different hosts to reuse the same configuration documents in slightly different ways. For example, the Bio2RDF Web Application R2 profile has been configured to resolve queries that include the new

[14] https://github.com/bio2rdf/bio2rdf-mapping

‘_resource’ and ‘_vocabulary’ namespaces (section 3.2.1), as well existing query types used by the base Bio2RDF profile, and to resolve these queries using the R2 SPARQL endpoints.

The Bio2RDF Web Application accepts RDF requests in the Accept Request and does not use URL suffixes for Content Negotiation, as most Linked Data providers do, as that would make it difficult to reliably distinguish identifiers across all of the namespaces that are resolved by Bio2RDF. Specifically, there is no guarantee that a namespace will not contain identifiers ending in the same suffix as a file format. For example, if a namespace had the identifier “plants.html”, the Bio2RDF Web Application would not be able to resolve the URI consistently to non-HTML formats using Content Negotiation. For this reason, the Bio2RDF Web Application directive to resolve HTML is a prefixed path, which is easy for any scriptable User Agent to generate. In the example above the identifier could be resolved to an RDF/XML document using “/rdfxml/namespace:plants.html’’, without any ambiguity as to the meaning of the request, as the file format is stripped from the prefix by the web application, based on the web application configuration.

3.2.9 Resolving Bio2RDF IRIs using multiple SPARQL endpoints

The Bio2RDF Web Application is designed to be used as an interface to a range of different Linked Data providers. It includes declarative rules that are used to map queries between the Bio2RDF IRI format and the identifiers used by each Linked Data provider. For example, the Bio2RDF R2 Web Application has been configured to resolve queries of the form

http://bio2rdf.org/uniprot:P05067

using UniProt's new SPARQL endpoint, currently available at http://beta.sparql.uniprot.org/sparql. In this way, as it becomes increasingly commonplace for data providers to publish their data at their own SPARQL endpoints, Bio2RDF will be able to leverage these resources and incorporate them into the Bio2RDF network, while still supporting queries that follow Bio2RDF IRI conventions.

3.3 Results

3.3.1 Bio2RDF Release 2

Nineteen datasets, including 5 new datasets, were generated as part of R2 (Table 2). R2 also includes 3 datasets that are themselves aggregates of datasets which are now available as one resource. For instance, iRefIndex consists of 13 datasets (BIND, BioGRID, CORUM, DIP, HPRD, InnateDB, IntAct, MatrixDB, MINT, MPact, MPIDB, MPPI and OPHID) while NCBO's BioPortal collection currently consists of 100 OBO ontologies including ChEBI, Protein Ontology and the Gene Ontology. We also have 10 additional updated scripts that are currently generating updated datasets and SPARQL endpoints to be available with the next release: ChemBL, DBPedia, GenBank, PathwayCommons, the RCSB Protein Databank, PubChem, PubMed, RefSeq, UniProt (including UniRef and UniParc) and UniSTS.

Dataset SPARQL endpoints are available at http://[namespace].bio2rdf.org. For example, the Saccharomyces Genome Database (SGD) SPARQL endpoint is available at http://sgd.bio2rdf.org. All updated Bio2RDF Linked Data and their corresponding Virtuoso DB files are available for download at http://download.bio2rdf.org.

Table 2 Bio2RDF Release 2 datasets with select dataset metrics. The asterisks indicate datasets that are new to Bio2RDF.

Dataset	Namespace	# of triples	# of unique subjects	# of unique predicates	# of unique objects
Affymetrix	affymetrix	44469611	1370219	79	13097194
Biomodels*	biomodels	589753	87671	38	209005
Comparative Toxicogenomics Database	ctd	141845167	12840989	27	13347992
DrugBank	drugbank	1121468	172084	75	526976
NCBI Gene	ncbigene	394026267	12543449	60	121538103
Gene Ontology Annotations	goa	80028873	4710165	28	19924391
HUGO Gene Nomenclature Committee	hgnc	836060	37320	63	519628
Homologene	homologene	1281881	43605	17	1011783
InterPro*	interpro	999031	23794	34	211346
iProClass	iproclass	211365460	11680053	29	97484111
iRefIndex	irefindex	31042135	1933717	32	4276466
Medical Subject Headings	mesh	4172230	232573	60	1405919
National Center for Biomedical Ontology*	ncbo	15384622	4425342	191	7668644
National Drug Code Directory*	ndc	17814216	301654	30	650650
Online Mendelian Inheritance in Man	omim	1848729	205821	61	1305149
Pharmacogenomics Knowledge Base	pharmgkb	37949275	5157921	43	10852303
SABIO-RK*	sabiork	2618288	393157	41	797554
Saccharomyces Genome Database	sgd	5551009	725694	62	1175694
NCBI Taxonomy	taxon	17814216	965020	33	2467675
Total	**19**	**1010758291**	**57850248**	**1003**	**298470583**

3.3.2 Metrics informed querying

Dataset metrics (section 3.2.5) provide an overview of the contents of a dataset and can be used to guide the development of SPARQL queries. Table 3 shows values for the type-relation-type metric in the DrugBank dataset. In the first row we note that 11,512 unique pharmaceuticals are paired with 56 different units using the 'form' predicate, indicating the enormous number of possible formulations. Further in the list, we see that 1,074 unique drugs are involved in 10,891 drug-drug interactions, most of these arising from FDA drug product labels.

Table 3 Selected DrugBank dataset metrics describing the frequencies of type-relation-type occurrences. The namespace for subject types, predicates, and object types is http://bio2rdf.org/drugbank_vocabulary.

Subject Type	Subject Count	Predicate	Object Type	Object Count
Pharmaceutical	11512	Form	Unit	56
Drug-Transporter-Interaction	1440	Drug	Drug	534
Drug-Transporter-Interaction	1440	transporter	Target	88
Drug	1266	Dosage	Dosage	230
Patent	1255	Country	Country	2
Drug	1127	product	Pharmaceutical	11512
Drug	1074	ddi-interactor-in	Drug-Drug-Interaction	10891
Drug	532	Patent	Patent	1255
Drug	277	mixture	Mixture	3317
Dosage	230	Route	Route	42
Drug-Target-Interaction	84	Target	Target	43

The type-relation-type metric gives the necessary information to understand how object types are related to one another in the RDF graph. It can also inform the construction of an immediately useful SPARQL query, without losing time generating 'exploratory'

queries to become familiar with the dataset model. For instance, the above table suggests that in order to retrieve the targets that are involved in drug-target interactions, one should specify the 'target' predicate, to link to a target from its drug-target interaction(s):

```
PREFIX drugbank_vocabulary:
<http://bio2rdf.org/drugbank_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?dti ?target ?targetName
WHERE {
        ?dti a drugbank_vocabulary:Drug-Target-Interaction .
        ?dti drugbank_vocabulary:target ?target .
        ?target rdfs:label ?targetName.
}
```

Some of the results of this query are listed in Table 4.

Table 4. Partial results from a query to obtain drug-target interactions from the Bio2RDF DrugBank SPARQL endpoint.

Drug Target Interaction IRI	Target IRI	Target label
drugbank_resource: DB00002_1102	drugbank_target:1102	"Low affinity immunoglobulin gamma Fc region receptor III-B [drugbank_target:1102]"@en
drugbank_resource: DB00002_3814	drugbank_target:3814	"Complement C1r subcomponent [drugbank_target:3814]"@en
drugbank_resource: DB00002_3815	drugbank_target:3815	"Complement C1q subcomponent subunit A [drugbank_target:3815]"@en
drugbank_resource: DB00002_3820	drugbank_target:3820	"Low affinity immunoglobulin gamma Fc region receptor II-b [drugbank_target:3820]"@en
drugbank_resource: DB00002_3821	drugbank_target:3821	"Low affinity immunoglobulin gamma Fc region receptor II-c [drugbank_target:3821]"@en

Dataset metrics can also facilitate federated queries over multiple Bio2RDF endpoints in a similar manner. For example, the following query retrieves all biochemical reactions

from the Bio2RDF BioModels endpoint that are kinds of “protein catabolic process”, as defined by the Gene Ontology in the NCBO BioPortal endpoint:

```
PREFIX biopax_vocab: <http://bio2rdf.org/biopax_vocabulary:>
SELECT ?go ?label count(distinct ?x)
WHERE {
        ?go rdfs:label ?label .
        ?go rdfs:subClassOf ?goparent OPTION (TRANSITIVE) .
        ?goparent rdfs:label ?parentlabel .
        FILTER strstarts(str(?parentlabel), "protein catabolic process")
        SERVICE <http://biomodels.bio2rdf.org/sparql> {
                ?x biopax_vocabulary:identical-to ?go .
                ?x a <http://www.biopax.org/release/biopax-
                level3.owl#BiochemicalReaction> .
        }
}
```

3.3.3 Bio2RDF dataset vocabulary-SIO mapping

The mappings between Bio2RDF dataset vocabularies and SIO make it possible to formulate queries that can be applied across all Bio2RDF SPARQL endpoints, and can be used to integrate data from multiple sources, as opposed to a priori formulation of dataset specific queries against targeted endpoints. For instance, we can ask for chemicals that effect the ‘Diabetes II mellitus’ pathway and that are available in tablet form using the Comparative Toxicogenomics Database (CTD) and the National Drug Codes (NDC) Bio2RDF datasets, and the mappings of their vocabularies to SIO:

```
define input:inference "http://bio2rdf.org/sio_mappings"
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX ctd_vocab: <http:bio2rdf.org/ctd_vocabulary:>
PREFIX ndc_vocab: <http://bio2rdf.org/ndc_vocabulary:>
SELECT ?chemical ?chemicalLabel
WHERE {
        #SIO_01126: ‘chemical substance'
        ?chemical a sio:SIO_011126.
        ?chemical rdfs:label ?chemicalLabel .
```

```
        #affects Diabetes mellitus pathway
        ?chemical ctd_vocab:pathway <http://bio2rdf.org/kegg:04930> .
        #dosage form: tablet, extended release
        ?chemical ndc_vocab:dosage-form
        <http://bio2rdf.org/ndc_vocabulary:00426c812b33febc3f9cd1fee8cc83ce> .
}
```

This query is possible because the classes 'ctd_vocab:Chemical' and 'ndc_vocab:human-prescription-drug' have been mapped as subclasses of the SIO class 'chemical substance'[15].

3.4 Discussion

Bio2RDF Release 2 marks several important milestones for the open source Bio2RDF project. First, the consolidation of scripts into a single GitHub repository will make it easier for the community to report problems, contribute code fixes, or contribute new scripts to add more data into the Bio2RDF network of linked data for the life sciences. Already, we are working with members of the W3C Linking Open Drug Data (LODD) to add their code to this GitHub repository, identify and select an open source license, and improve the linking of Bio2RDF data. With new RDF generation guidelines and example queries that demonstrate use of dataset metrics and provenance, we believe that Bio2RDF has the potential to become a central meeting point for developing the biomedical semantic web. Indeed, we welcome those that think Bio2RDF could be useful to their projects to contact us on the mailing list and participate in improving this community resource.

A major aspect of what makes Bio2RDF successful from a Linked Data perspective is the use of a central registry of datasets in order to normalize generated

[15] http://semanticscience.org/resource/SIO_011126

IRIs. Although we previously created a large aggregated namespace directory, the lack of extensive curation meant that the directory contained significant overlap and omissions. Importantly, no script specifically made use of this registry, and thus adherence to the namespaces was strictly in the hands of developers at the time of writing the code. In consolidating the scripts, we found significant divergence in the use of a preferred namespace for generating Bio2RDF IRIs, either because of the overlap in directory content, or in the community adopting another preferred prefix. With the addition of an API to automatically generate the preferred Bio2RDF IRI from any number of dataset prefixes (community-preferred synonyms can be recorded), all Bio2RDF IRIs can be validated such that unknown dataset prefixes must be defined in the registry. Importantly, our registry has been shared with maintainers of identifiers.org in order for their contents to be incorporated into the MIRIAM registry [81] which powers that URL resolving service. Once we have merged our resource listings, we expect to make direct use of the MIRIAM registry to list new entries, and to have identifiers.org list Bio2RDF as a resolver for most of its entries. Moreover, since the MIRIAM registry describes regular expressions that specify the identifier pattern, Bio2RDF scripts will be able to check whether an identifier is valid for a given namespace, thereby improving the quality of data produced by Bio2RDF scripts.

The dataset metrics that we now compute for each Bio2RDF dataset have significant value for users and providers. First, users can get fast and easy access to basic dataset metrics (number of triples, *etc.*) as well as more sophisticated summaries such as which types are in the dataset and how are they connected to one another. This data graph summary is the basis for SparQLed, an open source tool to assist in query composition

through context-sensitive autocomplete functionality. Use of these summaries also reduces the server load for data provider servers, which in turns frees up resources to more quickly respond to interesting domain-specific queries. Second, we anticipate that these metrics may be useful in monitoring dataset flux. Bio2RDF now plans to provide bi-annual release of data, and as such, we will develop infrastructure to monitor change in order to understand which datasets are evolving, and how are they changing. Thus, users will be better able to focus in on content changes and providers will be able to make informed decisions about the hardware and software resources required to provision the data to Bio2RDF users.

Our demonstration of using SIO to map Bio2RDF dataset vocabularies helps facilitate the composition of queries for the basic kinds of data or their relationships. Since SIO contains unified and rich axiomatic descriptions of its classes and properties, in the future we intend to explore how these can be automatically reasoned about to improve query answering with newly entailed facts as well as to check the consistency of Bio2RDF linked data itself.

4 Chapter: Ovopub - Modular data publication with minimal provenance

Abstract

With the growth of the Semantic Web as a medium for creating, consuming, mashing up and republishing data, our ability to trace any statement(s) back to their origin is becoming ever more important. Several approaches have now been proposed to associate statements with provenance, with multiple applications in data publication, attribution and argumentation. Here, we describe the ovopub, a modular model for data publication that enables encapsulation, aggregation, integrity checking, and selective-source query answering. We describe the ovopub RDF specification, key design patterns and their application in the publication and provenance of data in the life sciences.

Contribution to thesis

This chapter describes a model for linked data provenance that HyQue uses to automatically record basic provenance details for hypothesis evaluations, an important element of the HyQue framework (Objective #1). The ovopub is simple enough that implementation in HyQue is straightforward and can be extended if additional provenance details are needed. The application of the ovopub for capturing dataset-specific provenance details, such as the experimental method used to detect a protein-protein interaction (described in detail in this chapter, with example queries) also demonstrates how provenance information may be generated and subsequently used for hypothesis evaluation by HyQue data retrieval and evaluation functions (described in Chapter 5).

4.1 Data publication and attribution: A simple problem with a (so far) complicated solution

With its standards and ever-evolving best practices, the Semantic Web enables a virtuous cycle of data creation and data consumption in which consumers can also act as creators with the ability to re-mix, re-phrase and re-publish content. However, these activities raise important questions about the provenance of any given data: What is it? Who made it? Where did it come from? How was it made? When was it made? What license governs its use, and can I reuse, modify or sell it? From the scientific research perspective, facilitating data re-use with attribution incentivizes the publishing of quality datasets that include provenance [82]. Indeed, our trust of data depends on being able to uncover and assess its provenance.

The Web of Data initiative is focused on developing practices to support the unique naming of individual resources on the Web [83] and to describe their provenance [84, 85][16]. Important questions remain, however: how do we name, describe, publish and refer to assertions? Several efforts to address this challenge have so far been concerned with shrinking the scale of attributable and publishable objects from entire documents to some subset thereof. Nanopublications were developed to describe minimal assertions [86] between concepts, defined by [87] as the "smallest, unambiguous unit of thought". In its original formulation, a nanopublication consisted of two RDF named graphs: one graph containing a statement (an RDF triple) and another containing the annotations about that statement. Building on nanopublications, microattributions [88] aim to enable data attribution by specifying the source of statements included in data resources, and

[16] http://www.w3.org/TR/prov-overview/

have been used in describing gene variant data [89]. The nanopublications model has been recently updated [90] to use three named graphs: one for the statement(s), one for supporting statement(s) and one for related provenance. While certainly appealing on the surface, it is unclear what constitutes "support" in a nanopublication and whether supporting graphs can be exactly another's existing assertion graph in a different context. Similarly, while the micropublication model [91] uses a graph-based formalism of variable size and structure to construct an argumentation network linking textual statements and data as evidence for claims, it is unclear how a statement differs from a claim. These contributions are an important step forward, but do not address the larger, more inclusive question: How can we describe self-contained units of knowledge, of *any* size or level of complexity, such that they can be published, shared, reused, extracted, modified, and republished? Moreover, neither model has yet been demonstrated for selective fact-based information retrieval across potentially billions of similar statements in which the structure and irregularity of content may yet pose significant challenges.

Here we propose the ovopub as a data and knowledge publication model for describing any set or sequence of statements along with their provenance. We use ovo - from the Latin *ab ovo* - to refer to the earliest possible point in time at which an assertion could be described. The ovopub is simple by design and may be applied to represent statements and datasets of varying complexity and size. We posit that the ovopub is sufficient to describe any kind of statement and make it publishable and attributable, while its simple structure will facilitate widespread deployment to create, link, share and query ovopub networks. Specifically, the ovopub (i) embodies the simplest structure necessary to describe data, its provenance and digital rights, (ii) enables construction of

more complex statements and arguments, (iii) allows encapsulation of selected statements, (iv) facilitates source restricted information retrieval and (v) enables integrity checking of published data.

4.2 The ovopub: Linking statements with provenance

An ovopub *contains* and *links* to one or more statements about resources, including those describing its provenance (Figure 5). A resource is an object identified by a unique identifier. A statement is an n-tuple that either a) describes a relationship between two resources or b) assigns a literal value to a resource. There are two basic kinds of ovopubs: assertions and collections. An *assertion ovopub* contains one or more statements that may be true or false. A *collection ovopub* contains one or more assertions and/or ovopubs, and is specifically meant to share or restrict a search to the resources contained therein. Essential ovopub provenance includes the creator(s) of the ovopub, a timestamp of when the ovopub was created, and a license to specify the rights and responsibilities of the creator and user of the ovopub. The integrity of any ovopub may be validated against a computed hash that is associated with the content of, but external to, the ovopub [92] .

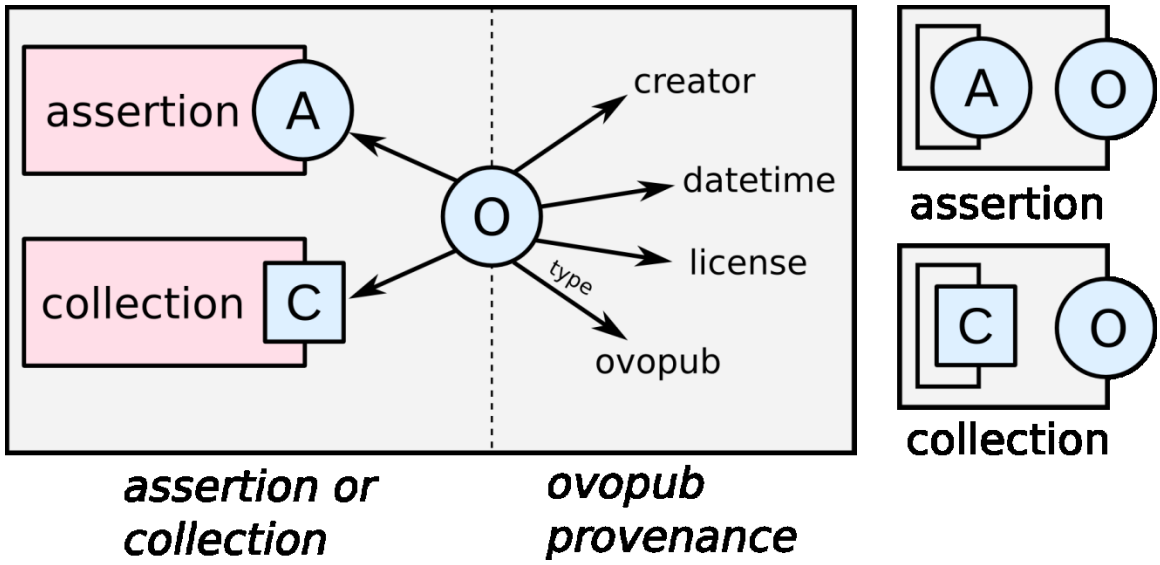


Figure 5 The basic structure of an ovopub.

An ovopub 'O' contains an assertion 'A' or collection 'C' of assertions and/or ovopubs, and links to its own provenance (right). We use two symbols (far right) throughout to distinguish the assertion ovopub and collection ovopub.

In section 4.2.1 we show how ovopubs may be chained together to describe relationships between assertions and also how ovopubs may be aggregated and the provenance of this aggregation described. In section 4.2.2 we present an RDF specification for ovopubs.

4.2.1 Patterns for building and extending ovopub networks

4.2.1.1 The Chaining Pattern

Ovopubs may be chained together through new assertion ovopubs in a manner that both retains the original provenance of each ovopub and describes the provenance of the new assertion(s). Figure 6 illustrates such a case where three resources E, F and G are linked together in an assertion ovopub O3. Assertion ovopub O1 contains an assertion A that describes the relationship P1 between E and F, and assertion ovopub O2 contains an assertion B describes the relationship P2 between F and G. Each ovopub may have been

created by different sources and/or at different times, and this information can be captured in their respective ovopub-linked provenance. Assertion ovopub O3 that contains an assertion C is semantically equivalent to collection ovopub O4, which contains a collection D of ovopubs O1 and O2, by virtue of the fact that it describes the same set of relations (P1 and P2, respectively) between resources E, F and G. Assertion ovopub O5 contains an assertion E stating that assertion A is *related-to* assertion C (we use *related-to* as a generic example predicate; any appropriate predicate may be used). Importantly, a collection ovopub like ovopub O4 may be invoked in order to isolate and specifically refer to a set of selected resources or ovopubs extracted from a potentially large and complex network.

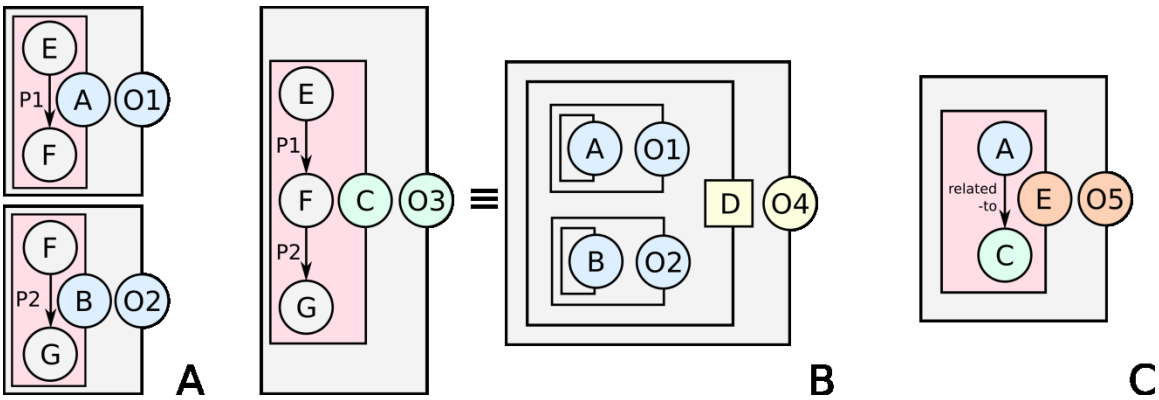


Figure 6 Chaining statements together using a combination of assertion and collection ovopubs. An assertion ovopub O3 contains the assertion C linking resources E, F and G. These statements can be described in individual assertions A and B in ovopubs O1 and O2 respectively, which are contained in collection D of ovopub O4. Assertion C in ovopub O3 is semantically equivalent to collection D in ovopub O4 because they describe the same relations, P1 and P2, between resources E, F and G. Assertion ovopub O5 contains an assertion E stating that assertion A is *related-to* assertion C.

The chaining pattern has innumerable potential uses, including but not limited to (i) providing a rich historical provenance of an object or assertion [93-95] (*e.g.* assertion

A *has-source* assertion C), (ii) establishing argumentation [91] (*e.g.* assertion A *supported-by*/*disputed-by* assertion C) and (iii) making ontological assertions [96] (*e.g.* axiom A *equivalent-to* axiom C). These specific cases differ from the general use case only by the predicate linking assertions A and C - the ovopub graph structures remain the same. In every case, the provenance for any ovopub is captured as part of the ovopub itself. However, should the ovopub creator decide to include explicit domain-specific provenance of the statement itself, another ovopub could link the first ovopub to its source provenance via an appropriate provenance relation.

4.2.1.2 The aggregation pattern

Given the potential for redundancy of assertions in an ovopub network, it becomes necessary to aggregate statements based on identity in the non-provenance content of an ovopub, or some other criteria of identity or similarity. Figure 7 shows how ovopubs can be used to aggregate assertions. An assertion ovopub O describes an assertion A. Assertion ovopubs O1, O2 and O3 describe related assertions A1, A2 and A3, respectively. Assertion ovopub O4 describes the aggregation relationship between A, A1, A2 and A3, in a new assertion B. Finally, ovopub O5 groups all assertion ovopubs together into collection C. The ovopub aggregation pattern is broader than that of the nanopublication cardinal assertion, where aggregation can only occur over those assertions that contain exactly the same subject, predicate and object [87, 97].

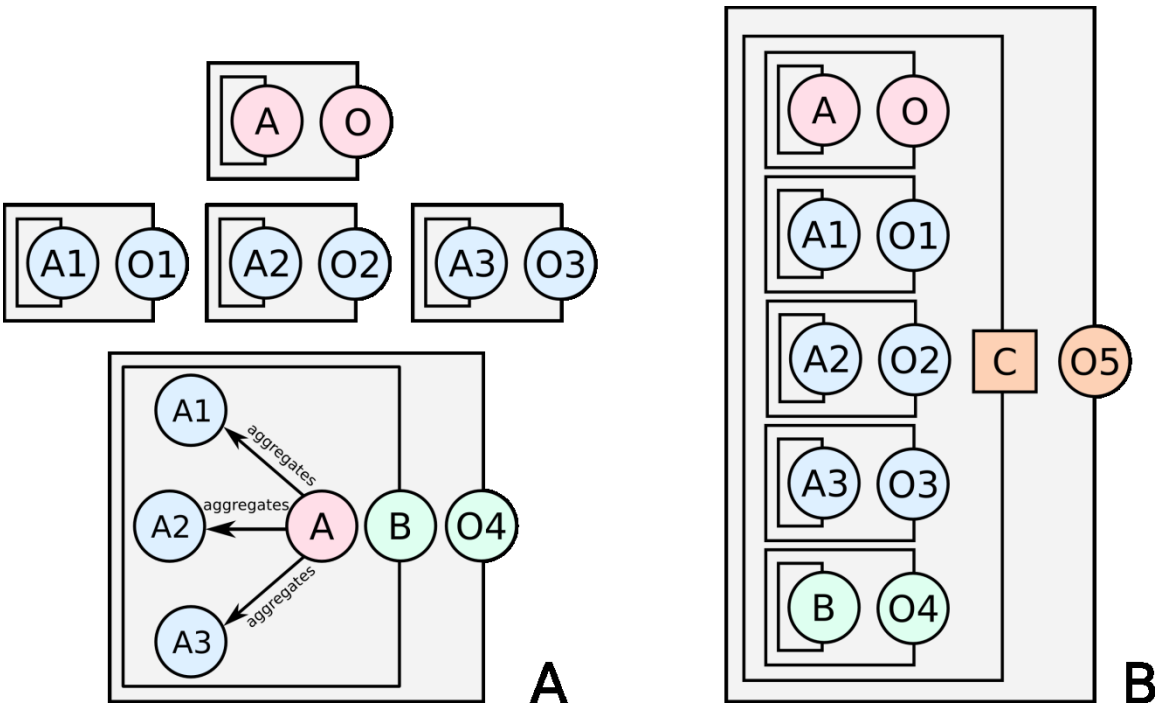


Figure 7 Aggregating statements into a collection ovopub.

A single assertion ovopub O4 describes an aggregation relationship between assertion A in ovopub O and three assertions A1, A2 and A3 individually described in ovopubs O1, O2 and O3. All assertion ovopubs can be grouped into collection C described in ovopub O5.

4.2.2 Ovopub RDF specification

Ovopubs that are represented using the Resource Description Framework (RDF) must implement the following specification in order to be considered a valid ovopub. An ovopub is a named graph that contains an assertion graph of statements of interest, as a named sub-graph. Thus, all statements must be expressed as quads such that the referent graph is the assertion URI (for statements in the assertion graph) or ovopub URI (for statements in the ovopub). We use five vocabularies for the necessary types and relations: RDF (rdf) [98], RDFG (rdfg) [99], RDF schema (rdfs) [100], XML Schema (xsd) [101], Dublin Core Metadata Terms (dc) [102], and the Semanticscience Integrated Ontology

(sio) [103]. An assertion ovopub (Figure 8) must be typed as sio:assertion-ovopub[17] and its assertion sub-graph must be explicitly linked to the ovopub using rdfg:subGraphOf. A collection ovopub must be typed as a sio:collection-ovopub[18] and its collection sub-graph must explicitly linked to the ovopub, also using rdfg:subGraphOf. All ovopubs and their sub-graphs must have unique identifiers. The sio:assertion-ovopub and sio:collection-ovopub are both subclasses of sio:ovopub[19]. Ovopubs may optionally use rdfs:label to specify a title (with appropriate language tag), dc:identifier to specify a source specified non-URI identifier, and dc:description to provide a more detailed description (with language tag). The ovopub creator must be specified using dc:creator linked to either a literal specified as an xsd:string or to a resolvable linked data URI (*e.g.* a FOAF entry). The ovopub date of creation must be specified as a timestamp using dc:date as the datatype property and the literal value a xsd:dateTime. The ovopub license must be specified using dc:rights with the value specified as a URI pointing to the license document. We recommend using the Creative Commons by attribution (CC-BY) licenses for data.

[17] http://semanticscience.org/resource/SIO_001302
[18] http://semanticscience.org/resource/SIO_001301
[19] http://semanticscience.org/resource/SIO_001300

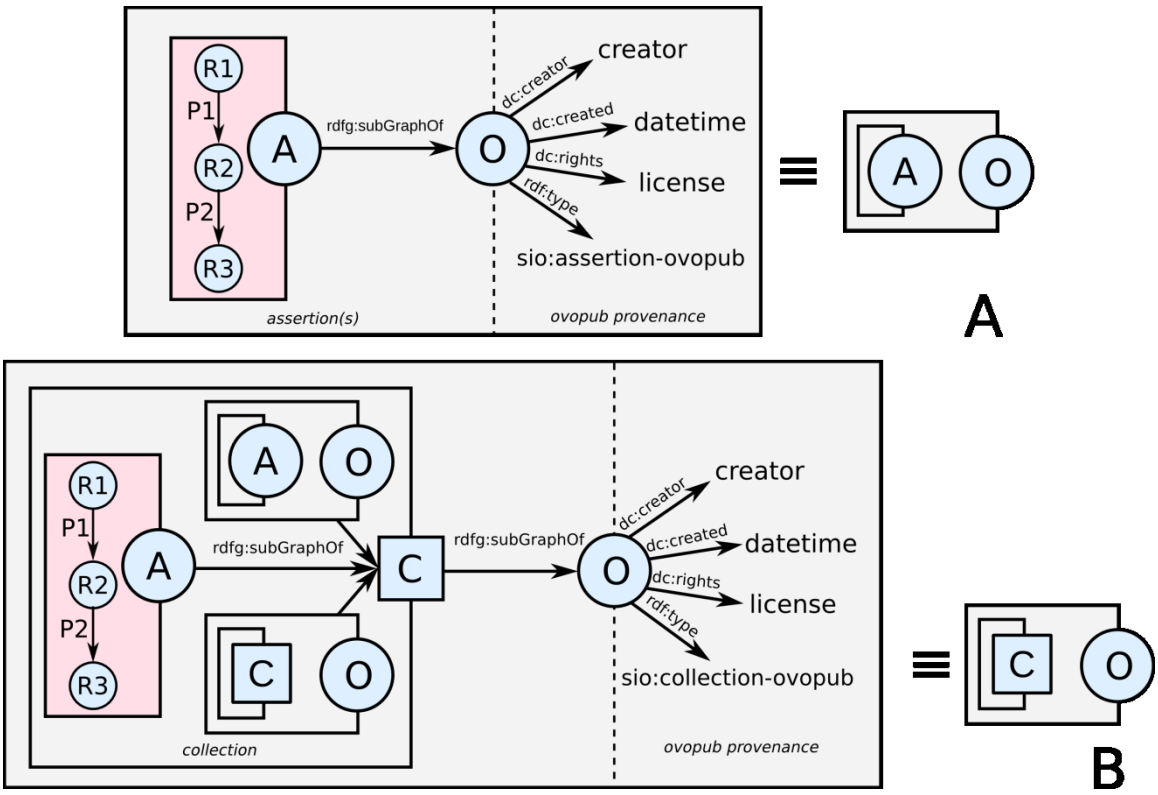


Figure 8 Assertion (A) and collection (B) ovopub RDF specifications.

(A) An assertion ovopub O links to a named sub-graph A containing one or more statements using rdfg:subGraphOf. (B) A collection ovopub O links to a named-subgraph C containing one or more statements or ovopubs. Each ovopub type is specified using the rdf:type relation. Provenance details for both assertion and collection ovopubs are described using relations from the Dublin Core (DC) vocabulary.

4.3 Modular knowledge representation with ovopubs

In this section we illustrate the application of the ovopub model to a complex life sciences data scenario: publishing the iRefIndex database of protein-protein interactions (PPIs). iRefIndex [104] collects interactions reported in 13 source databases and groups interactions based on taxon and sequence identity as well as sequence similarity of molecular interaction participants. In this fashion, iRefIndex serves as a natural aggregation point for protein and PPI data from multiple sources and must therefore

distinguish the relations that were asserted by the source databases and the relations that iRefIndex asserts for the purpose of data aggregation. Figure 9 illustrates (i) capturing the relations between an individual PPI (sourced from BioGRID:464511) and the dataset of which it is a part, (ii) the description of interactions in terms of their protein participants, method of PPI identification and publication, and (iii) the aggregation of PPIs into interaction groups based on the aggregation criteria set out by iRefIndex.

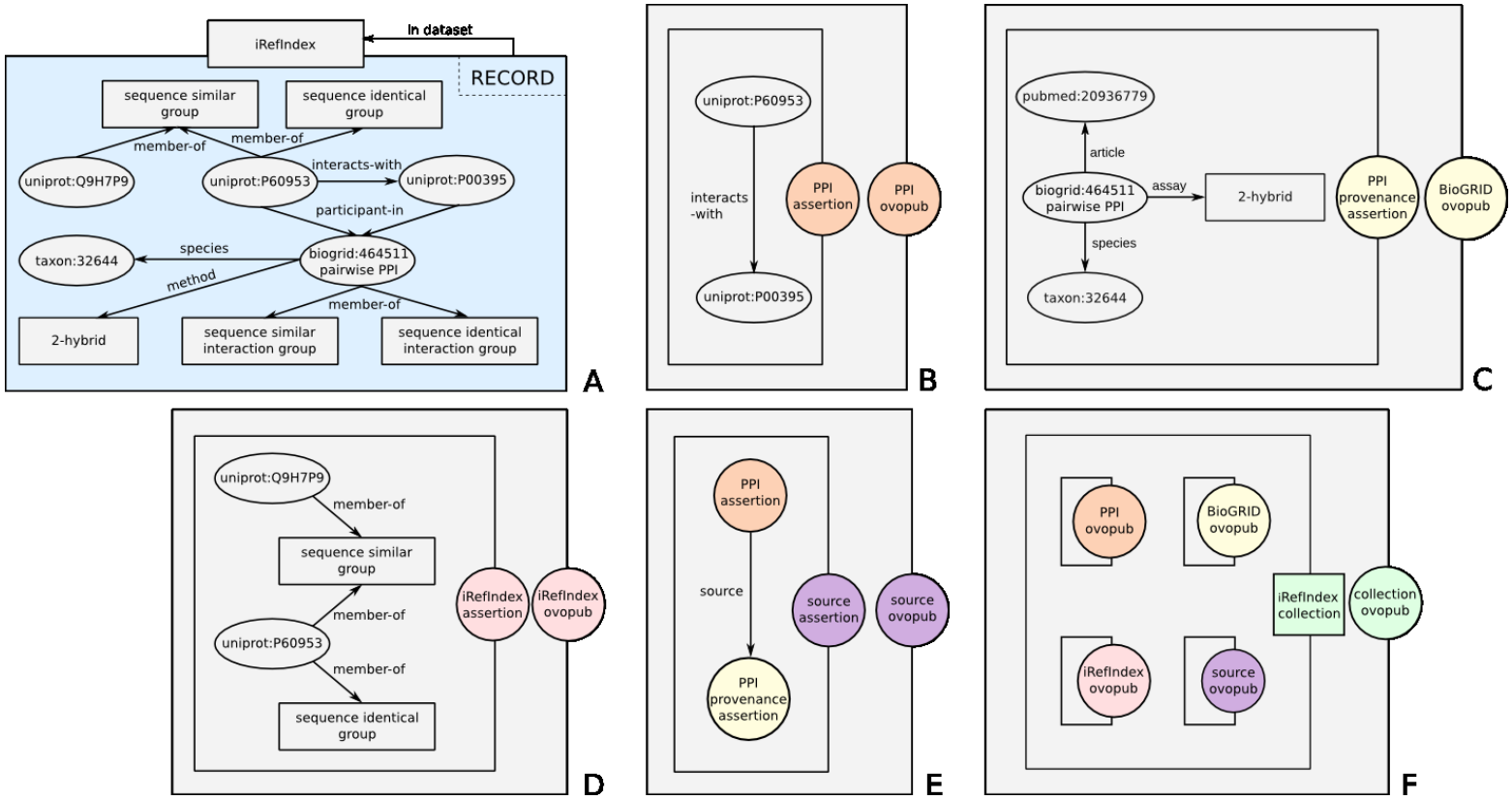


Figure 9 Relations between data items in an iRefIndex record for BioGRID:464511 (A) and their corresponding representation as ovopubs.

An assertion ovopub (B) describes the protein-protein interaction (PPI) relationship between two UniProt proteins. An assertion ovopub (C) describes the provenance of this PPI as collected by BioGRID, while an assertion ovopub (D,) describes the membership of the UniProt proteins in iRefIndex sequence similarity groups. An assertion ovopub (E) links the PPI assertion to the BioGRID provenance assertion, to indicate the PPI source. A collection ovopub (F) collects the assertion ovopubs.

This demonstrates the utility of ovopubs for (i) encapsulation and versioning, (ii) description of domain knowledge and (iii) aggregation. Ovopubs, and thus their provenance, can be as fine-grained as the simplest statement in a dataset. For example, the simplest ovopub in this case is the ovopub containing an assertion describing the protein interaction tuple. The provenance for this PPI from BioGRID is individually described it its own ovopub, as are the iRefindex generated sequence groupings for the PPI participants, and the link between the PPI assertion and the BioGRID provenance assertion. Lastly, these individual ovopubs are grouped together in an iRefindex collection ovopub.

Based on the model illustrated in Figure 9, we generated ovopubs for the iRefIndex Bio2RDF dataset, sourced from Release 10 of iRefIndex[20]. Metrics for this dataset are listed in Table 5.

Table 5 Metrics for iRefIndex ovopub dataset

# of triples	69 852 574
# of distinct subjects	5 221 353
# of distinct predicates	32
# of distinct objects	3 883 725
# of distinct literals	16 983 725
# of ovopubs	4 111 051
# of assertions	2 990 250

4.4 Using ovopubs for context-sensitive information retrieval

To illustrate how ovopubs can be used to describe and retrieve context-specific provenance details, and how ovopubs are linked to their content, we present 5 natural language questions about PPIs described in the iRefIndex dataset and their ovopubs, as

[20] http://irefindex.org/download/irefindex/data/archive/release_10.0/psi_mitab/MITAB2.6/

well as the corresponding SPARQL queries. For each query we also present the query results.

Query 1 demonstrates how to retrieve ovopubs that describe assertions about specific PPIs. Query 2 demonstrates how to retrieve the provenance associated with specific ovopubs. Query 3 demonstrates how to retrieve iRefIndex-specific provenance associated with a specific PPI, independent of its ovopub provenance. Query 4 demonstrates how to retrieve ovopub-specific provenance details by restricting the content of an assertion graph within the ovopub. Lastly, Query 5 demonstrates how to retrieve content across the databases compiled by iRefIndex in order to assess the reporting prevalence of PPIs.

Query 1: In which ovopub(s) is serine/threonine-protein kinase PAK 1 (O88643) reported to interact with G25K GTP-binding protein (P60766)?

```
PREFIX vocab: <http://bio2rdf.org/irefindex_vocabulary:>
PREFIX rdfg: <http://www.w3.org/2004/03/trix/rdfg-1/>
SELECT DISTINCT ?ovopub ?label
WHERE {
GRAPH ?assertion {
?interaction vocab:interactor_a <http://bio2rdf.org/uniprot:O88643> .
?interaction vocab:interactor_b <http://bio2rdf.org/uniprot:P60766> .
}
?assertion rdfg:subGraphOf ?ovopub .
?ovopub rdfs:label ?label .
}
```

This query returns 8 ovopubs (one for each PPI assertion), each with an identifier as well as a human-readable label. The results are shown in Table 6.

Table 6 Results of Query 1 for ovopubs describing a PPI between uniprot:O88643 and uniprot:P60766.

Ovopub	Label
ovopub_c58ca2d7386820d2160693d2919f4145	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by affinity chromatography technology (biogrid:502946)
ovopub_efa4ec07960344e899c1635aaa5ec09a	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by affinity chromatography technology (bind:315278)
ovopub_c4aeb19b61c3b3d79e075089c413aa0a	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by pull down (dip:DIP-70378E)
ovopub_df692a8f2e4adbb1e9918a9ff0672e57	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by two hybrid (intact:EBI-651254)
ovopub_3fe1751f954d1188bc10fc286b53c9b2	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by two hybrid (intact:EBI-651243)
ovopub_0f577d2097d0335194941d4273026cbe	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by pull down (mint:MINT-1791686)
ovopub_bec06712015be511f33c27c0d2f790ab	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by two hybrid (mint:MINT-19832)
ovopub_63d4ee87d63da1f1280b1a7fd0ab4364	Assertion ovopub for Pairwise interaction between uniprotkb:O88643 and uniprotkb:P60766 identified by two hybrid (mint:MINT-19631)

Query 2: What provenance is associated with the ovopub describing the interaction between serine/threonine-protein kinase PAK 1 (O88643) and serine/threonine-protein kinase OSR1 (O95747)?

```
PREFIX vocab: <http://bio2rdf.org/irefindex_vocabulary:>
PREFIX rdfg: <http://www.w3.org/2004/03/trix/rdfg-1/>
PREFIX dc: <http://purl.org/dc/terms/>
SELECT ?creator ?date ?rights ?type
WHERE {
GRAPH ?assertion {
?interaction vocab:interactor_a <http://bio2rdf.org/uniprot:O88643> .
?interaction vocab:interactor_b <http://bio2rdf.org/uniprot:O95747> .
}
?assertion rdfg:subGraphOf ?ovopub .
GRAPH ?ovopub {
?ovopub dc:created ?date  .
?ovopub dc:rights ?rights .
?ovopub dc:creator ?creator .
?ovopub rdf:type ?type .
}
}
```

This query returns the following:

Date: 2014-04-04T16:44:19Z

Rights: "check source for further restrictions"

Creator: A Callahan

Type: assertion ovopub (SIO_001302)

Query 3: What is the iRefIndex-specific provenance associated with the interaction between serine/threonine-protein kinase PAK 1 (O88643) and serine/threonine-protein kinase OSR1 (O95747)?

```
PREFIX vocab: <http://bio2rdf.org/irefindex_vocabulary:>
PREFIX np: <http://www.nanopub.org/nschema#>
SELECT ?s ?p ?o
WHERE {
GRAPH ?assertion {
        ?interaction vocab:interactor_a <http://bio2rdf.org/uniprot:O88643> .
        ?interaction vocab:interactor_b <http://bio2rdf.org/uniprot:O95747> .
}
?assertion np:hasProvenance ?assertion_provenance .
GRAPH ?assertion_provenance {
        ?s ?p ?o .
}
}
```

This query returns annotations from iRefIndex that describe the provenance of the PPI, including the source article (PMID:14707132), the minimum number of interactions reported (4), and the source organism (*S. cerevisiae*, TAXID:4932). The complete results are shown in Table 7.

Query 4: Who are the creators of ovopubs that involve serine/threonine-protein kinase PAK 1 (O88643) as an interactor?

```
PREFIX vocab: <http://bio2rdf.org/irefindex_vocabulary:>
PREFIX rdfg: <http://www.w3.org/2004/03/trix/rdfg-1/>
PREFIX dc: <http://purl.org/dc/terms/>
SELECT DISTINCT ?creator
WHERE {
GRAPH ?assertion {
        ?interaction vocab:interactor_a <http://bio2rdf.org/uniprot:O88643> .
}
?assertion rdfg:subGraphOf ?ovopub .
?ovopub dc:creator ?creator .
}
```

This query returns the creators of ovopubs describing specific PPIs. In this case, the query returns only one result – "A Callahan".

Table 7 Results of Query 3 for the iRefIndex-sourced provenance for a specific PPI.

Subject	Predicate	Object
http://bio2rdf.org/intact:EBI-620898	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://bio2rdf.org/mi:0915
http://bio2rdf.org/intact:EBI-620898	http://purl.org/dc/terms/created	2013-03-02T00:00:00Z
http://bio2rdf.org/intact:EBI-620898	http://bio2rdf.org/irefindex_vocabulary:article	http://bio2rdf.org/pubmed:14707132
http://bio2rdf.org/intact:EBI-620898	http://bio2rdf.org/irefindex_vocabulary:number-supporting-articles	"1"^^<http://www.w3.org/2001/XMLSchema#int>
http://bio2rdf.org/intact:EBI-620898	http://bio2rdf.org/irefindex_vocabulary:maximum-number-interactions-reported	"4"^^<http://www.w3.org/2001/XMLSchema#int>
http://bio2rdf.org/intact:EBI-620898	http://bio2rdf.org/irefindex_vocabulary:minimum-number-interactions-reported	"4"^^<http://www.w3.org/2001/XMLSchema#int>
http://bio2rdf.org/intact:EBI-620898	http://bio2rdf.org/irefindex_vocabulary:source	http://bio2rdf.org/mi:0469
http://bio2rdf.org/intact:EBI-620898	http://bio2rdf.org/irefindex_vocabulary:expansion-method	"none"@en
http://bio2rdf.org/intact:EBI-620898	http://bio2rdf.org/irefindex_vocabulary:host-organism	http://bio2rdf.org/taxonomy:4932

Query 5: How many databases report the top 5 most reported interactions?

```
PREFIX vocab: <http://bio2rdf.org/irefindex_vocabulary:>
SELECT DISTINCT ?int_a ?int_b COUNT(?int) AS ?num_interactions COUNT(DISTINCT
?db) AS ?num_dbs
WHERE {
?int vocab:interactor_a ?int_a .
?int vocab:interactor_b ?int_b .
?int vocab:source-database ?db .
} ORDER BY DESC(COUNT(?int)) LIMIT 5
```

This query counts the number of interaction reports for protein pairs in iRefIndex, as well as the databases that report them. It demonstrates that ovopub networks can be queried for the content of their assertions, without requiring any reference to ovopub structure or content. The results of this query are shown in Table 8.

Table 8 Results of Query 4 for the most reported PPIs, the number of reports, and the number of reporting databases.

Interactor A (Uniprot ID)	Interactor B (UniProt ID)	# of interaction reports	# of reporting databases
FACT complex subunit SPT16 (P32558)	FACT complex subunit POB3 (Q04636)	118	6
Eukaryotic translation initiation factor 3 subunit B (P06103)	Eukaryotic translation initiation factor 3 subunit A (P38249)	111	5
Histone H3 (P61830)	Histone H4 (P02309)	108	7
UBA domain-containing protein RUP1 (Q12242)	Ubiquitin carboxyl-terminal hydrolase 2 (Q01476)	89	4
Inosine-5'-monophosphate dehydrogenase 4 (P50094)	Inosine-5'-monophosphate dehydrogenase 3 (P50095)	85	3

All of the top five most reported PPIs were observed in yeast (*Saccharomyces cerevisiae*). The most frequently reported interaction is between SPT16 (P32558) and POB3 (Q04636), which are known to form a stable heterodimer [105] as part of the FACT complex. The FACT complex is a chromatin factor crucial in nucleosome reorganization during DNA replication and transcription [105], and its function is conserved across species including humans [106]. The second most reported interaction is between translation initiation factors eIF3B (P06103) and eIF3A (P38249). These proteins make up part of the translation initiation factor 3 (eIF3) core complex, which is a component of the molecular machinery involved in protein synthesis [107]. The third most reported interaction is between histones H3 (P61830) and H4 (P02309), proteins which form a core part of nucleosomes and also the upstream activation factor (UAF) that stimulates transcription [108]. The fourth most reported interaction is between UBA domain-containing protein RUP1 (Q12242) and ubiquitin carboxyl-terminal hydrolase 2

(Q01476). These proteins form a complex that is an antagonist and potential regulator of the HECT ubiquitin-protein ligase RSP5, an essential ubiquitin ligase that plays a role in various protein trafficking functions within and into the cell [109]. Lastly, the fifth most reported interaction is between IMP dehydrogenases 4 (P50094) and 5 (P50095), thought to form heteromers [110] that, like their homotetrameric counterparts [111], catalyze the first rate-limiting step of guanine nucleotide biosynthesis *in vivo* and have thus been studied as potential targets for chemotherapy drugs. Considering all of these PPIs together, it is not surprising that they constitute the 5 most reported interactions in iRefIndex. They all occur in *S. cerevisiae*, one of the most popular model organisms, and all describe interactions that facilitate and in some cases regulate crucial (and thus well-studied) biological functions - nucleotide biosynthesis, DNA replication, transcription, protein synthesis and protein trafficking.

For each interacting pair, it is also possible to ask *which* of the databases in iRefIndex report the interaction. For example, for the most frequently occurring pair the necessary query is:

```
PREFIX vocab: <http://bio2rdf.org/irefindex_vocabulary:>
SELECT DISTINCT ?db
WHERE {
?int vocab:interactor_a <http://bio2rdf.org/uniprot:P32558> .
?int vocab:interactor_b <http://bio2rdf.org/uniprot:Q04636> .
?int vocab:source-database ?db .
}
```

The results of this query are: BioGrid, BIND, MINT, DIP, IntAct and MIPS.

4.5 Discussion

The ovopub is simple by design in order to maximize its reuse in scenarios with differing source data and use requirements. The ovopub is a single object that contains and links to

basic elements of data-centric provenance (what, who, when, rights) and its content, whether a simple assertion involving objects, other assertions or a collection ovopub, or a collection of ovopub assertions and/or collections. The ovopub model enables encapsulation, linking and aggregation and facilitates complex queries that consider elements of provenance and trust.

The ovopub can be readily contrasted with the nanopublication and micropublication models. The assertion ovopub is simpler as it consists of a named graph containing a single assertion graph, with key provenance information directly contained in and associated with the ovopub graph. In contrast, the nanopublication graph is explicitly linked by three relations to three named graphs: one that contains the statement(s), one that contains the supporting statement(s) and one that contains the related provenance. The ovopub therefore reduces the number of required statements by consolidating the statement and provenance graph into a single ovopub graph and leaving the supporting graph to be specified as another ovopub. It also removes the ambiguity of what should be included in the supporting statement graph, whether an assertion ovopub can be used as a supporting graph in another publication, and sidesteps the problem of how to manage a change in the supporting statement graph *vis-a-vis* the original nanopublication. In addition, unlike the nanopublication or the micropublication, the provenance of an ovopub is distinct from the provenance of the ovopub payload, whose historical provenance (who stated it? how was it generated? where was the statement obtained? *etc*.) could either be published a set of provenance-oriented assertion ovopubs or could be directly stated in the ovopub provenance.

Cardinal assertions aggregate syntactically identical statements described in nanopublications to establish a confidence or evidence score [97]. Through the iRefIndex exemplar, we show that ovopubs can be used to aggregate syntactically identical, semantically equivalent, or semantically similar statements from multiple sources (each being described in an ovopub) from which evidential strengths may be computed if desired. Cardinal assertions can be easily computed through hash sums on the payload assertion or the specified members of the dataset as identified through relations provided by explicit reification. More importantly, the integrity of an ovopub can also be assessed using an implementation of trusty URIs [92]. In this case, the entirety of the ovopub would be subject to a hash sum whose value could be recorded in a separate ovopub signed by the creator. Data 'hijacking' by adding additional links or statements to an ovopub after it has been created and published could be detected by virtue of the fact that the hash of their content would differ from that of the original ovopub.

With the availability of increasingly powerful triple stores and reasoners, adopting the ovopub model for existing large linked data networks is certainly feasible. OpenLink Virtuoso and Ontotext's OWLIM provide cluster-capable triple store implementations, while the Digital Enterprise Research Institute at the National University of Ireland Galway with Fujitsu Laboratories have announced a new linked data platform capable of querying billions of triples at greatly improved speeds [112]. Similarly, the Web-scale Parallel Inference Engine (WebPIE) enables large scale reasoning [113].

In summary, the ovopub is a new model for data publication. Its simplicity and modular design support the creation of networks of arbitrary size and complexity. We have described the minimal elements required for an ovopub, as well as how ovopubs can

be used to make assertions about objects, literals, statements or even collections of statements. We have described the application of ovopubs to address the requirements of description, encapsulation, aggregation, data integrity and demonstrated selective-source query answering over a life science dataset. Our next steps are to implement the ovopub model for Bio2RDF datasets, evaluate aspects related to performance and scalability, and to explore the use of ovopub networks for knowledge discovery.

5 Chapter: Data integration and reasoning on the Semantic Web to identify aging-related genes in *C. elegans*

Abstract

Background: Extensive studies have been carried out on *C. elegans* as a model organism to elucidate mechanisms of aging and the effects of perturbing known aging-related genes on lifespan and behavior. This research has generated large amounts of experimental data that is increasingly difficult to manually integrate and analyze with existing databases and domain knowledge. There is a need for tools and methods to assist the biologist in evaluating hypotheses about aging that take advantage of the growing amounts of experimental data and the knowledge captured in bio-ontologies.

Results: Using HyQue, a Semantic Web tool for hypothesis-based querying and evaluation, we evaluated 48,231 *C. elegans* genes for their role in aging. We also contributed three new datasets to the Bio2RDF network of linked data for the life sciences – WormBase, GenAge and GenDR. By evaluating data retrieved from these resources and integrated with existing Bio2RDF datasets, HyQue correctly identified 8 genes known to regulate lifespan in *C. elegans*. HyQue also identified 24 candidate genes whose effects on lifespan and aging have not been well characterized.

Conclusions: We demonstrate a scalable approach to semi-automated hypothesis evaluation by applying HyQue to hypotheses about the role of *C. elegans* genes in aging and lifespan, using existing experimental data integrated with domain ontologies. In doing so we identified candidate genes whose effects on lifespan are not well understood, several of which have human orthologs with known effects on aging-related diseases including Parkinson's and Alzheimer's disease. HyQue, as well as the functions and rules

it used to evaluate *C. elegans* genes for their roles in aging and lifespan are publicly available and can be extended and repurposed to evaluate additional hypotheses.

Contribution to thesis

In this chapter, I describe the application of HyQue to evaluating hypotheses about the role of genes in aging and lifespan in *C. elegans*, to achieve Objective #2 of this thesis – using HyQue to evaluate biological hypotheses. This work is an important extension of Chapter 2 in two ways: firstly, unlike hypotheses about galactose metabolism in yeast, which is a well-understood system with extensive experimental validation (thus making it a good candidate for a first 'test run' application of HyQue), which genes govern aging and longevity in *C. elegans* is an evolving area of investigation. Using HyQue to investigate this domain is an important contribution to the field of aging research. Secondly, the application of HyQue described in this chapter is at a much larger scale that that previously described – we used HyQue to evaluate 48,231 gene-central hypotheses and analyze the results of this process. While Chapter 2 described a proof-of-concept application of HyQue over a small number of hypotheses, the work described in this Chapter represents a real-world use case for HyQue with biologically significant results, including findings verified by independent data.

5.1 Introduction

The biology of aging is a significant area of biomedical research, motivated by a desire to uncover the mechanisms that govern aging and to regulate these processes towards developing effective therapies for age-related diseases. Experiments using model organisms have identified genes, gene variations, and biological pathways that regulate lifespan [114] in humans [115, 116] and model organisms such as the nematode *C. elegans* [117]. Mutations in many of the genes responsible for regulating longevity in model organisms are also implicated in human disease [118, 119]. Environmental factors such as dietary restriction [120], temperature [121, 122] and pheromone exposure [123] have also been found to have significant effects on lifespan in model organisms, often acting through stress response genes and pathways whose activity are triggered by changes in nutrient availability [114].

Biologists studying the role of genes in aging use a variety of approaches, and a typical experiment involves perturbing environmental conditions or gene expression *in vivo*, measuring changes in lifespan (both temporal and reproductive), and measuring associated changes in gene expression to identify potential agents mediating any observed lifespan effects. The use of high-throughput experimental techniques such as microarrays and next generation sequencing systems capable of measuring changes in expression of thousands of genes, combined with the large body of existing experimental data, publications and databases dedicated to capturing aging-associated annotations makes it increasingly intractable for scientists to manually sift through these resources in the context of their own research. Large-scale bioinformatics analyses of genes and aging pathways have seen recent success in identifying candidate aging-related genes and

elucidating their effects on biological pathways [124-127], and the fruits of this labour are increasingly accessible by the scientific community. For example, the Human Aging Genomics Resources group maintains the GenAge and GenDR databases of both model organism and human genes and their experimentally determined effects on lifespan and aging. Model organism databases such as WormBase also capture gene and phenotype annotations related to longevity and lifespan. Such resources will continue to grow in both size and span as more experimental data is generated.

With the increased availability of data about aging-related phenomena, a significant challenge lies in finding, integrating, and evaluating this information to address questions of biological interest, in this case in discovering genes that are responsible for aging and lifespan. Semantic Web technologies including the Web Ontology Language (OWL) and the Resource Description Framework (RDF) can enable such applications by allowing scientists to represent data and knowledge in a machine understandable way, such that we can leverage computational power to query and reason over them [128, 129]. More specifically, life sciences data on the Semantic Web such as Bio2RDF [16] and the growing number of bio-ontologies enable data integration and powerful question answering in a variety of biological and biomedical domains [32, 130, 131]. Motivated by these developments, we use HyQue [132, 133] to evaluate hypotheses on the roles of *C. elegans* genes in aging. HyQue is a Semantic Web tool that uses W3C standards (RDF/OWL) for representing data, domain-specific knowledge and evaluation rules to computationally evaluate biological hypotheses using existing data and knowledge resources. Specifically, we developed aging domain-specific hypothesis evaluation rules and used HyQue to execute them in a single pass over all *C. elegans*

genes. We demonstrate that HyQue is a scalable, semantic approach to uncover new candidate aging-related genes and assign a measure of confidence to the role of previously characterized aging-associated genes.

5.2 Methods

5.2.1 HyQue system overview and architecture

HyQue [133] is a rule-based system that retrieves and evaluates evidence relevant to a hypothesis. HyQue rules are specified using SPIN [134], which is based on SPARQL, a W3C standard query language for RDF. In the following sections, we describe the HyQue Ontology for hypotheses, events, and hypothesis evaluations, design patterns for rules, data retrieval and data evaluation functions, and explain how HyQue uses these functions to calculate aging-specific event and hypothesis scores. Figure 10 provides an overview of the HyQue system. HyQue takes as input a hypothesis specified in RDF, and a set of domain specific SPIN rules. It executes the SPIN rules to retrieve relevant RDF data and OWL ontologies, and evaluates the evidence obtained. HyQue generates an RDF output that includes the evaluation, the rules used, and references to the supporting data.

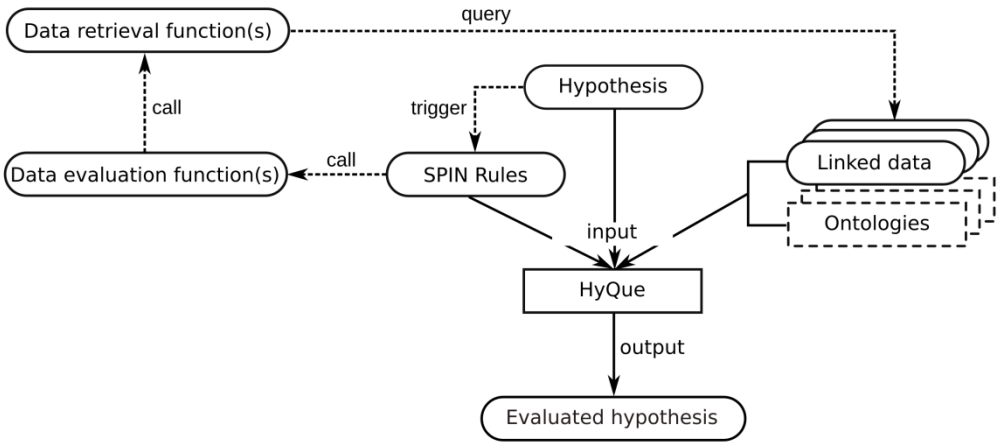


Figure 10 HyQue system architecture.

HyQue takes as input a hypothesis, SPIN rules, and a knowledge base composed of data and ontologies. The hypothesis triggers SPIN rules that retrieves and evaluates relevant data. HyQue produces as output an evaluation that scores the hypothesis, and provides links to the hypothesis, SPIN rules, and data used.

5.2.2 HyQue ontology for hypotheses, events, and evaluations

HyQue uses the HyQue Hypothesis Ontology (HO)[21] to specify a valid hypothesis and its evaluation. A HyQue *hypothesis* is composed of one or more *events* that are related to each other through one or more *propositions.* An event is a process that involves one or more participants (*e.g.* agents, targets), while a proposition connects two or more events through logical operators AND, OR, and XOR. The operators control how the overall hypothesis score is calculated (described below). A *hypothesis evaluation* includes a score for the hypothesis and each of the components of the hypothesis (propositions, events) and their provenance. We use the ovopub (Chapter 4) model for associating the provenance of the evaluation (hypothesis, SPIN rules, evidence). An ovopub is a graph

[21] http://semanticscience.org/ontology/hyque.owl

that contains an assertion composed of one or more triples along with the metadata of the ovopub (type, creator, creation time, and license).

5.2.3 Design patterns for HyQue functions and rules

HyQue uses two kinds of rules to evaluate a hypothesized event – **domain specific rules** that are triggered by the biological event type (*e.g.* 'gene induction', 'aging') and **system rules** which are triggered by the output of the domain specific rules and the operators that relate events in the hypothesis to calculate an overall hypothesis score. All HyQue domain rules are linked with the generic HyQue event class (**hyque:HYPOTHESIS_0000004**) and then filtered to domain specific event types in the **WHERE** clause of the rule.

HyQue domain rules consist of data evaluation functions and data retrieval functions. A data retrieval function executes a SPARQL query over a specified data source to obtain data about an entity of interest (typically a hypothesis event participant). A data evaluation function evaluates the result of a data retrieval function in the context of the biological domain associated with the hypothesized event. Specifically, data evaluation functions call one or more data retrieval functions, assess the retrieved data, and then return a Boolean or numeric value to quantify the assessment. Event scores are calculated by aggregating the output of data evaluation functions into a single evaluation score. Data evaluation and event scoring functions are combined in a SPIN rule associated with specific event type(s).

HyQue system rules automatically generate proposition and overall hypothesis evaluation scores from individual event scores generated by the domain rules. These scores are calculated in a bottom-up procedure, in which first event scores are calculated,

followed by the proposition scores, and finally the overall hypothesis score. For a proposition that specifies events related by the AND operator, HyQue calculates the proposition score by taking the mean of the individual event scores. For a proposition that specifies events related by an OR operator, HyQue takes the maximum event score as the proposition score. For a proposition that describes a single event (with the XOR operator) the event score is assigned as the proposition score. This procedure is iteratively repeated to calculate the overall hypothesis score. As each score is calculated HyQue generates statements linking the score to the function(s) used to calculate the score, thereby ensuring that provenance of each part of HyQue's evaluation process is recorded in the evaluation itself.

5.2.4 Integrating experimental data and annotations about aging in *C. elegans*

HyQue evaluates the role of C. elegans genes in aging using a variety of data sources including existing curated databases and raw data, terminologies and ontologies. In the following sections, we describe how we prepared each of the seven data sources for use in HyQue, and related data analysis and transformation processes used.

5.2.4.1 Linking aging data on the Semantic Web

A number of databases dedicated to cataloguing genes that regulate the biological processes of aging have recently been developed, including the GenAge and GenDR databases developed by the Human Ageing Genomic Resources (HAGR) group [135] and the human-curated WormBase database [136]. GenAge describes genes that are known to affect longevity and aging [135], while GenDR describes genes that confer lifespan extension under dietary restriction or whose expression is found to be significantly different under dietary restriction across multiple studies [126]. WormBase

annotates *C. elegans* genes with genetic and protein sequence data, known phenotypes and their roles in biological pathways, including those specific to aging processes, as well as links to the literature. It is not possible to automatically query across these independently maintained resources to collect all the information they contain about a given gene. Thus, to enable integration of these databases as evidence sources for HyQue, we generated linked data [12] versions of each. We used the Resource Description Framework (RDF) [137] and Bio2RDF best practices [138] to facilitate dataset interoperability and querying. At the core of the Bio2RDF approach is the use of uniform resource identifiers (URIs) for consistently naming entities and the relationships that hold between them. Using the Bio2RDF approach ensures that an entity is automatically assigned the same identifier in every dataset that contains statements about it, such that queries across multiple datasets using the same identifier (query federation) will retrieve all statements about a given entity.

5.2.4.2 Linked Open Data relevant to the biology of aging

Relevant sources of data currently available in Bio2RDF (Release 2) include release 9.0 of the iRefIndex database of experimentally determined protein-protein interactions (PPIs) [104], and the Gene Ontology Annotations (GOA) database of protein function, process and cellular location annotations [24] processed in 2012. As described below, these datasets are used in concert with the linked data versions of GenAge, GenDR and WormBase by HyQue to retrieve data about PPIs and functional annotations for *C. elegans* genes.

5.2.4.3 Gene expression data and analysis

Next-generation sequencing technologies (NGS) measure system-wide gene expression changes under varying experimental conditions. We searched the NCBI's Gene Expression Omnibus (GEO) database for datasets from experiments that targeted biochemical pathways in *C. elegans* related to aging, to allow HyQue to retrieve and evaluate gene expression data. We identified two relevant RNA-seq datasets - GSE39574 and GSE36041. The GSE39574 dataset quantifies changes in gene expression when the transcription factor unc-62 (known to regulate lifespan and aging [139]) is knocked down. The GSE36041 dataset contains the expression profiles of three *C. elegans* models with impaired IGF-1 signaling (the IGF-1 signaling pathway is a well-characterized regulator of longevity [140]). We identified significant differences in gene expression of variants *versus* control(s) using TopHat and CuffLinks as described in [141].

To integrate our genomic data analysis results with the Bio2RDF linked data resources described above, we developed a data model to represent the RNA-seq data analysis results as linked data. Our model describes experiments and experimental conditions, samples, and the resulting gene expression and gene expression change values, as well as the relations that hold between them. It re-uses WormBase and Gene Ontology identifiers for genes and phenotypes where possible, and associates each gene expression change value with its corresponding statistical confidence value (p-value) as well as the gene expression values it is derived from.

5.2.4.4 Quantifying GO annotation co-occurrence

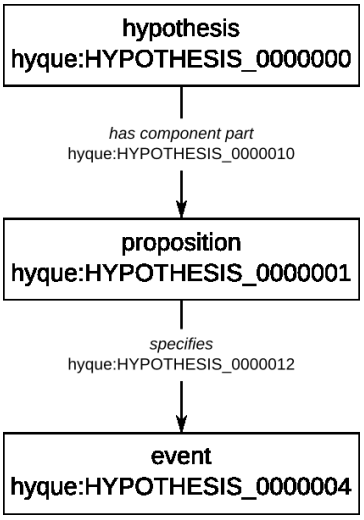
Co-occurrence frequencies of GO terms used for gene and gene product annotations have been used to discover and confirm associations between biological functions and

processes [142, 143]. In the context of evaluating the role of a given gene in aging, the co-occurrence of its GO annotations with GO terms related to aging is of interest. To measure these co-occurrences, we programmatically calculated the co-occurrence frequency of each pair of GO terms in the GOA database and generated linked data describing these frequencies.

5.2.5 Tailoring HyQue to the aging domain

To model the hypothesis that a *C. elegans* gene is involved in aging using HO, we used the Gene Ontology term 'aging' (GO:0007568) as the event type of interest, where an aging event has a gene as an agent specified with its WormBase Bio2RDF identifier (e.g. 'daf-2' is specified as 'http://bio2rdf.org/wormbase:WBGene00000898'). The RDF representation of such a hypothesis is shown in BOX 1. We developed domain specific data retrieval and evaluation functions (triggered by an event of type 'aging') for investigating the role of *C. elegans* genes in aging.

BOX 1 RDF of HyQue hypothesis that daf-2 is the agent in an aging event




```
@prefix hyp:
<http://bio2rdf.org/hyqueData:aging_hypothesis_> .
@prefix hyque:
<http://semanticscience.org/ontology/hyque.owl#> .
@prefix hybrow:
<http://semanticscience.org/ontology/hybrow.owl#> .

hyp:h rdf:type hyque:HYPOTHESIS_0000000 ;
   hyque:HYPOTHESIS_0000010 hyp:p0 .

hyp:p0 rdf:type hyque:HYPOTHESIS_0000014 ;
   hyque:HYPOTHESIS_0000012 hyp:e1 .

hyp:e1 a hyque:HYPOTHESIS_0000004 ;
   rdf:type <http://bio2rdf.org/go:0007568> ;
   hyque:HYPOTHESIS_0000017 "0";
   hyque:HYPOTHESIS_0000015
      <http://bio2rdf.org/wormbase:WBGene00000898> .
```

5.2.6 Building HyQue data retrieval functions

HyQue uses 14 data retrieval functions (each identified by "DRF" followed by a unique number) over 7 data sources to collect the evidence used to assess the involvement of a gene in aging. Each function is gene-centric in that it queries a data source to retrieve annotations associated with the *C. elegans* gene specified in the hypothesis. The 14 data retrieval functions (with a brief description *in italics*) are:

DRF1 Retrieve *C. elegans* gene product UniProt identifiers from GOA

The Bio2RDF UniProt GOA dataset maps gene symbols to UniProt protein identifiers. This DRF retrieves UniProt identifiers of the protein products of a C. elegans gene specified by its gene symbol.

DRF2 Retrieve GO term co-occurrence frequencies from GOA annotation term co-occurrences

We calculated the GO term co-occurrence frequencies for each pair of GO terms used in GOA. This DRF retrieves the frequency of co-occurrence of a given pair of GO terms.

DRF3 Retrieve *C. elegans* gene identifier and lifespan effect from GenAge

The GenAge Bio2RDF dataset we generated includes the assigned GenAge identifier and known lifespan effect(s), if any, for C. elegans genes. This DRF retrieves the GenAge identifier and lifespan effect of a given C. elegans gene, specified by its WormBase identifier.

DRF4 Retrieve *C. elegans* gene expression fold change and probability value from NGS GEO datasets

Our analysis of NGS GEO datasets GEO35974 and GEO36041 quantified gene fold change (relative to a control) and associated p-values from experimental data. This DRF

retrieves gene fold change values and associated p-values for a C. elegans gene specified by its WormBase identifier.

DRF5 Retrieve *C. elegans* gene GO annotation and evidence type from WormBase

The Bio2RDF WormBase dataset provides GO annotations and source evidence type (using the ECO ontology) for C. elegans genes. This DRF retrieves the GO annotation and associated evidence type for a C. elegans gene specified by its WormBase identifier.

DRF6 Retrieve *C. elegans* gene symbol from WormBase

WormBase provides the official gene symbols for C. elegans genes. This DRF retrieves the gene symbol for a C. elegans gene specified by its WormBase identifier.

DRF7 Retrieve *C. elegans* interacting genes with extended or shortened lifespan phenotype from WormBase

The Bio2RDF WormBase dataset includes curated genetic interactions between C. elegans genes. This DRF retrieves any interacting genes that have the WormBase 'extended lifespan' or 'shortened lifespan' phenotype for a C. elegans gene specified by its WormBase identifier.

DRF8 Retrieve gene product interacting proteins and interaction provenance from iRefIndex

The Bio2RDF iRefIndex dataset describes protein-protein interactions (PPIs), associated experimental methods, and number of publications that report a given interaction. This DRF retrieves PPIs involving proteins specified by their UniProt identifier.

DRF9 Retrieve *C. elegans* gene product GO process annotations from GOA

The Bio2RDF GOA dataset includes GO process annotations for proteins specified with UniProt identifiers. This DRF retrieves GO annotations for the protein products of a C.

elegans gene specified by its WormBase identifier where the UniProt identifiers of the protein products of the gene were retrieved by DRF1.

DRF10 Retrieve *C. elegans* gene associated phenotype from WormBase

The Bio2RDF WormBase dataset provides phenotypes associated with C. elegans genes. This DRF retrieves phenotypes associated with a C. elegans gene specified by its WormBase identifier.

DRF11 Retrieve *C. elegans* gene DR-associated gene expression change values from GenDR

The Bio2RDF GenDR dataset includes experimentally determined gene expression changes associated with dietary restriction (DR). This DRF retrieves gene expression change values under DR for a C. elegans gene specified by its gene symbol.

DRF12 Retrieve *C. elegans* gene product interacting proteins from iRefIndex

This DRF retrieves PPIs for the protein products (retrieved by DRF1) of a C. elegans gene specified by its WormBase identifier, as well as the experimental method used to detect each PPI and the number of reporting publications (retrieved by DRF8).

DRF13 Retrieve *C. elegans* gene-phenotype associations from GenDR

The Bio2RDF GenDR dataset contains DR-related gene-phenotype associations. This DRF retrieves the DR-related phenotypes for a C. elegans gene specified by its WormBase identifier.

DRF14 Retrieve *C. elegans* gene phenotypes induced by RNAi from WormBase

The Bio2RDF WormBase dataset includes the results of RNAi experiments, including the phenotypes associated with genes whose expression was diminished by RNAi. This DRF

retrieves the observed phenotypes from RNAi experiments targeting a given C. elegans gene specified by its WormBase identifier.

Below, we present the data retrieval functions **DRF3** and **DRF12** in detail and include an example result for each. **DRF3** retrieves a gene's approved gene symbol from the Bio2RDF WormBase dataset, and uses this symbol to query the Bio2RDF GenAge dataset for its effect on lifespan (BOX 2).

BOX 2 SPARQL query used for DRF3

```
PREFIX wormbase_vocabulary: <http://bio2rdf.org/wormbase_vocabulary:>
PREFIX genage_vocabulary: <http://bio2rdf.org/genage_vocabulary:>
SELECT ?genageGene ?effect
WHERE {
  SERVICE <http://wormbase.bio2rdf.org/sparql> {
    {
      ?gene wormbase_vocabulary:has_sequence/cosmid_name ?arg1 .
      ?gene wormbase_vocabulary:has_approved_gene_name ?name .
    }
    UNION
    {
      ?arg1 a wormbase_vocabulary:Gene .
      ?arg1 wormbase_vocabulary:has_approved_gene_name ?name .
    } .
  } .
  SERVICE <http://genage.bio2rdf.org/sparql> {
    ?genageGene a genage_vocabulary:ModelOrganismAgingRelatedGene .
    ?genageGene genage_vocabulary:gene-symbol ?name .
    ?genageGene genage_vocabulary:lifespan-effect ?effect .
  } .
}
```

Example results for gene sams-1 are shown in Table 9.

Table 9 Results of DRF3 to retrieve a gene's lifespan effect from GenAge

GenAge gene identifier	Lifespan effect
http://bio2rdf.org/genage:0584	"increase"

DRF12 (BOX 3) requires the coordination of several Bio2RDF data sets, and thus is composed of calls to three other data retrieval functions. Specifically, because iRefIndex uses UniProt identifiers to describe protein-protein interactions, this data retrieval function must retrieve the UniProt identifier for the protein products of a given *C. elegans* gene (specified using its WormBase identifier). To do this it first retrieves the gene symbol for a given WormBase gene identifier using **DRF6** (BOX 4). Using the gene symbol, HyQue then queries the Bio2RDF GOA dataset for the corresponding UniProt protein identifiers associated with the symbol (there may be more than one) using **DRF1** (BOX 5). The resulting UniProt identifiers are used to query iRefindex for interacting proteins, the experimental method used to detect the interaction and the number of articles reporting the interaction with **DRF8** (BOX 6). The results of DRF12 for the *C. elegans* gene sams-1 are shown in Table 10.

BOX 3 SPARQL query used for DRF12

```
SELECT ?otherProtein ?count ?method
WHERE {
   ( ?arg1 ) :DRF6 ( ?symbol ) .
   BIND (fn:concat("^", ?symbol, "$") AS ?symbolstring) .
   ( ?symbolstring ) :DRF1 ( ?protein ) .
   ( ?protein ) :DRF8 ( ?otherProtein ?count ?method ) .
}
```

BOX 4 SPARQL query used for DRF6

```
PREFIX wormbase_vocabulary: <http://bio2rdf.org/wormbase_vocabulary:>
SELECT ?symbol
WHERE {
  SERVICE <http://wormbase.bio2rdf.org/sparql> {
    ?arg1 wormbase_vocabulary:has_approved_gene_name ?symbol .
  } .
}
```

BOX 5 SPARQL query used for DRF1

```
PREFIX goa_vocabulary: <http://bio2rdf.org/goa_vocabulary:>
SELECT ?protein
WHERE {
  SERVICE <http://goa.bio2rdf.org/sparql> {
    ?protein goa_vocabulary:gene_symbol ?goasymbol .
    ?protein goa_vocabulary:taxid <http://bio2rdf.org/taxon:6239> .
    FILTER regex(?goasymbol, ?arg1) .
  } .
}
```

BOX 6 SPARQL query used for DRF8

```
PREFIX irefindex_vocabulary: <http://bio2rdf.org/irefindex_vocabulary:>
SELECT DISTINCT ?otherProtein ?articles ?method
WHERE {
  SERVICE <http://irefindex.bio2rdf.org/sparql> {
    {
      ?interaction irefindex_vocabulary:interactor_a ?arg1 .
      ?interaction irefindex_vocabulary:interactor_b ?otherProtein .
      ?interaction irefindex_vocabulary:number-supporting-articles ?articles .
      ?interaction irefindex_vocabulary:method> ?method .
    }
    UNION
    {
      ?interaction2 irefindex_vocabulary:interactor_a ?otherProtein .
      ?interaction2 irefindex_vocabulary:interactor_b ?arg1 .
      ?interaction2 irefindex_vocabulary:number-supporting-articles ?articles .
      ?interaction2 irefindex_vocabulary:method ?method .
    } .
  } .
}
```

Table 10 Results DRF12 to retrieve interacting proteins from iRefIndex

UniProt protein identifier	Number of supporting articles	Experimental method identifier
http://bio2rdf.org/uniprot:O17680	1	http://bio2rdf.org/psi-mi:0397
http://bio2rdf.org/uniprot:O17680	1	http://bio2rdf.org/psi-mi:0398
http://bio2rdf.org/uniprot:P48181	1	http://bio2rdf.org/psi-mi:0676
http://bio2rdf.org/uniprot:P48181	1	http://bio2rdf.org/psi-mi:0109
http://bio2rdf.org/uniprot:P50305	1	http://bio2rdf.org/psi-mi:0397
http://bio2rdf.org/uniprot:P50305	1	http://bio2rdf.org/psi-mi:0398
http://bio2rdf.org/uniprot:P50306	1	http://bio2rdf.org/psi-mi:0397
http://bio2rdf.org/uniprot:P50306	1	http://bio2rdf.org/psi-mi:0398
http://bio2rdf.org/uniprot:Q27522	1	http://bio2rdf.org/psi-mi:0397
http://bio2rdf.org/uniprot:Q27522	1	http://bio2rdf.org/psi-mi:0398

The complete SPIN RDF representation of all functions and rules is available at the HyQue SPIN Rule GitHub repository[22].

5.2.7 Building HyQue data evaluation functions

We developed 9 domain-specific data evaluation functions (each identified by "DEF" followed by a unique number) for HyQue, each dedicated to assessing different a type of retrieved evidence for its contribution to a gene's involvement in aging, and each evaluating data returned by one or more data retrieval functions. These data evaluation functions answer the following questions for a given gene:

DEF1 Does the gene have a human-curated aging- or longevity-associated annotation?

DEF2 Is the gene significantly differentially expressed (under- or over-expressed) when genes that regulate known aging-related pathways are manipulated?

DEF3 Is the gene or a mammalian homolog significantly differentially expressed under dietary restriction across multiple studies?

[22] https://github.com/alisoncallahan/hyque-spin-rules

DEF4 Is the gene's effect on life-span extension under dietary restriction altered when its expression is manipulated?

DEF5 Does the gene (or its knockdown) have the extended or shortened lifespan phenotype in WormBase?

DEF6 Does the gene have aging-related functional annotations, where the annotation is derived from experimental evidence?

DEF7 Does the gene encode a protein that interacts with other proteins with aging-related functional annotations?

DEF8 Does the gene interact with other genes that extend or shorten lifespan?

DEF9 Does the gene have functional annotations that co-occur with aging-related functional annotations?

Each HyQue data evaluation function has two forms: one that returns a Boolean (TRUE or FALSE) if the condition specified in the WHERE clause of the function is satisfied or not, and a second that converts this Boolean to a numeric score (*e.g.* 1 for TRUE or 0 for FALSE). The second form of each evaluation function is used to calculate a quantitative score for each hypothesized event. HyQue system rules process event scores calculated in this way to generate overall hypothesis scores (see below). Below, we describe the data evaluation functions DEF1 and DEF7 that call the data retrieval functions presented in the previous section, as well as example results.

DEF1 calls **DRF3**, the data retrieval function that queries the GenAge Bio2RDF endpoint for a gene's GenAge identifier and lifespan effect. It then processes the retrieved data, and returns TRUE if the lifespan effect retrieved for a gene is "increase",

and FALSE otherwise (BOX 7). For the sams-1 gene, DEF1 returns TRUE, which is converted to an evaluation score of 1 as described above.

BOX 7 SPARQL query used for DEF1

```
ASK WHERE {
   ( ?gene ) :DRF3 ( ?genageGene ?effect ) .
   FILTER (?effect = "increase") .
}
```

DEF7 calls **DRF12**, the data retrieval function that retrieves PPIs involving protein products of the gene of interest, as well as **DRF9** (not shown), which retrieves GO process annotations for interacting proteins. It processes the retrieved data and returns TRUE if the experimental method associated with the PPI is one of a set of high-confidence detection methods and if the interacting protein's GO annotation is related to aging processes, and FALSE otherwise (BOX 8). For the sams-1 gene, DEF1 returns FALSE, which is converted to an evaluation score of 0. A description of each of the PPI detection methods used in DEF7 is provided in Table 15 of Appendix A.

BOX 8 SPARQL query for DEF7

```
ASK WHERE {
   ( ?gene ) :DRF12 ( ?protein ?articles ?method ) .
   ( ?protein ) :DRF9 ( ?goTerm ) .
   FILTER (((((?goTerm = <http://bio2rdf.org/go:0007568>) || (?goTerm =
<http://bio2rdf.org/go:0007569>)) || (?goTerm = <http://bio2rdf.org/go:0035982>))
|| (?goTerm = <http://bio2rdf.org/go:0010259>)) || (?goTerm =
<http://bio2rdf.org/go:0008340>)) .
   FILTER ((((((((((((((((((((((((((((((((((((((((((((((?method = <http://bio2rdf.org/psi-mi:0007>) ||
(?method = <http://bio2rdf.org/psi-mi:0254>)) || (?method = <http://bio2rdf.org/psi-
mi:0004>)) || (?method = <http://bio2rdf.org/psi-mi:0676>)) || (?method =
<http://bio2rdf.org/psi-mi:0096>)) || (?method = <http://bio2rdf.org/psi-mi:0114>))
|| (?method = <http://bio2rdf.org/psi-mi:0006>)) || (?method =
<http://bio2rdf.org/psi-mi:0019>)) || (?method = <http://bio2rdf.org/psi-mi:0415>))
|| (?method = <http://bio2rdf.org/psi-mi:0424>)) || (?method =
<http://bio2rdf.org/psi-mi:0107>)) || (?method = <http://bio2rdf.org/psi-mi:0055>))
|| (?method = <http://bio2rdf.org/psi-mi:0405>)) || (?method =
<http://bio2rdf.org/psi-mi:0434>)) || (?method = <http://bio2rdf.org/psi-mi:0435>))
|| (?method = <http://bio2rdf.org/psi-mi:0423>)) || (?method =
<http://bio2rdf.org/psi-mi:0077>)) || (?method = <http://bio2rdf.org/psi-mi:0364>))
|| (?method = <http://bio2rdf.org/psi-mi:0411>)) || (?method =
<http://bio2rdf.org/psi-mi:0809>)) || (?method = <http://bio2rdf.org/psi-mi:0040>))
|| (?method = <http://bio2rdf.org/psi-mi:0109>)) || (?method =
<http://bio2rdf.org/psi-mi:0069>)) || (?method = <http://bio2rdf.org/psi-mi:0410>))
|| (?method = <http://bio2rdf.org/psi-mi:0067>)) || (?method =
<http://bio2rdf.org/psi-mi:0678>)) || (?method = <http://bio2rdf.org/psi-mi:0012>))
|| (?method = <http://bio2rdf.org/psi-mi:0020>)) || (?method =
<http://bio2rdf.org/psi-mi:0826>)) || (?method = <http://bio2rdf.org/psi-mi:0515>))
|| (?method = <http://bio2rdf.org/psi-mi:0841>)) || (?method =
<http://bio2rdf.org/psi-mi:0417>)) || (?method = <http://bio2rdf.org/psi-mi:0728>))
|| (?method = <http://bio2rdf.org/psi-mi:0406>)) || (?method =
<http://bio2rdf.org/psi-mi:0870>)) || (?method = <http://bio2rdf.org/psi-mi:0858>)) .
}
```

5.2.8 Evaluating *C. elegans* genes for their role in aging processes

Using the nine data evaluation functions described above, we developed an aging-specific rule triggered by the 'aging' (GO:0007568) event type that calls each of the functions, and calculates an overall score for the hypothesis that a given gene is involved in aging.

Conceptually, this HyQue rule is triggered by any hypothesis that a gene is involved in an aging event. The event score is calculated by computing the sum of the outputs of each of the nine data evaluation functions listed above, and dividing this value by the maximum possible score (in this case, 9). For example, a gene that received a score of 1 for 6 of the 9 data evaluation functions would have a normalized score of 6/9 or 0.67, while a gene that satisfied only 3/9 data evaluation functions would have a normalized score of 0.33. This resulting normalized event score is processed by HyQue system rules to automatically generate proposition and overall hypothesis scores, using the logical operators specified for propositions as described above.

We executed the aging rule and functions over each of the 48,231 genes identified in WormBase using a Java implementation of HyQue that uses Jena 2.6.11 and the SPIN API 1.2.1. Using a machine with 4GB of RAM, processing all 48,231 genes required approximately 48 hours of computing time.

5.3 Results

5.3.1 WormBase, GenAge, and GenDR Bio2RDF datasets

The WormBase Bio2RDF dataset (built from Release WS235) contains 20,016,596 statements (or triples) about 33 types (or classes) of entities, with 41 relations between those types[23]. In addition to its own native identifiers, the WormBase dataset uses Gene Ontology (GO) for process/function annotations and PubMed identifiers for publications. It also uses the Evidence Codes Ontology (ECO) to specify the type of evidence that is the source of *C. elegans* gene-GO associations. The GenAge Bio2RDF dataset contains

[23] See http://download.bio2rdf.org/release/3/wormbase/wormbase.html for detailed metrics

63,474 statements about 16 types, with 42 relations[24]. The GenAge dataset uses NCBI Gene, Ensembl, UniProt, NCBI Taxonomy and PubMed identifiers for genes, proteins, species and publications, respectively. The GenDR Bio2RDF dataset contains 11,081 statements about 15 types, with 34 relations[25]. The GenDR dataset uses NCBI Gene, WormBase, NCBI Taxonomy, and PubMed identifiers for genes, phenotypes, species, and publications, respectively. Figure 11 shows partial records in the Bio2RDF versions of WormBase, GenAge and GenDR for the gene sams-1.

[24] See http://download.bio2rdf.org/release/3/genage/genage.html
[25] See http://download.bio2rdf.org/release/3/gendr/gendr.html

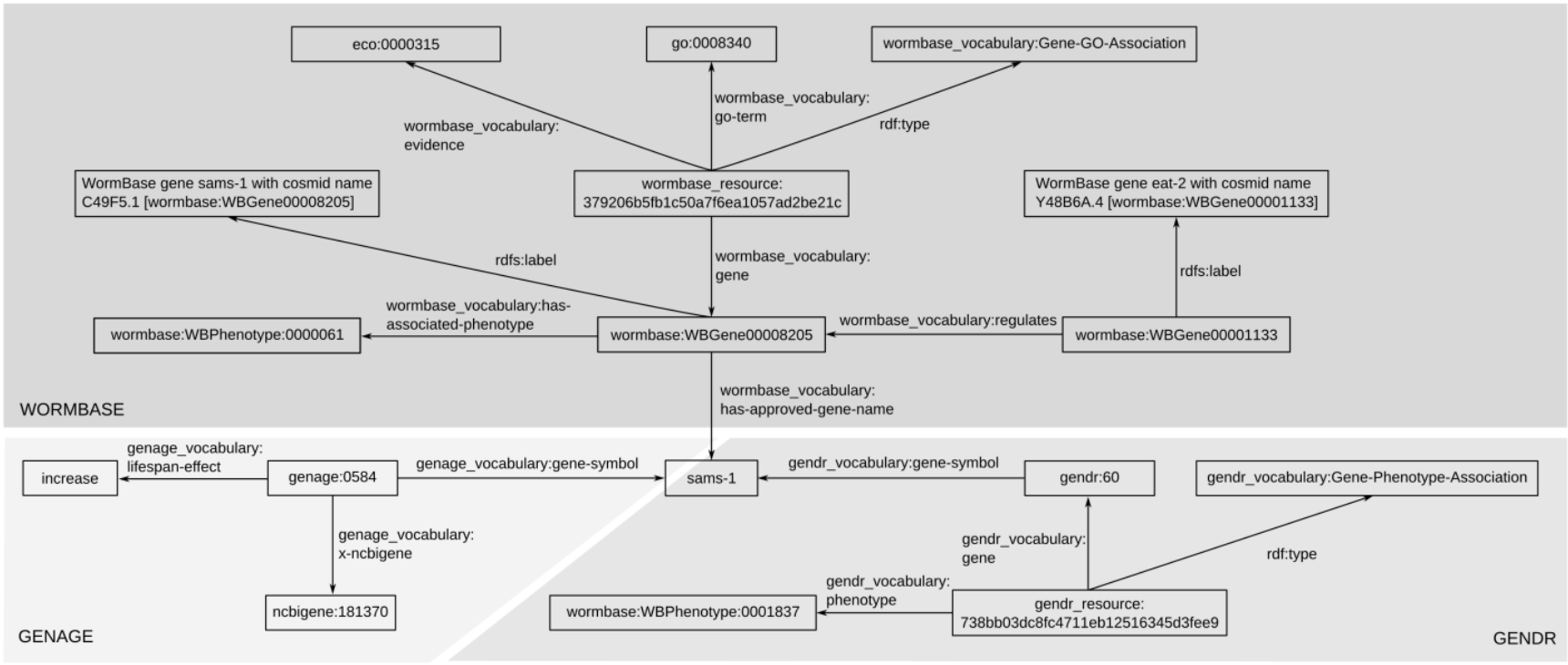


Figure 11 Information about the sams-1 gene in Bio2RDF Versions of WormBase, GenAge and GenDR.

WormBase includes phenotypes, GO associations, and regulatory relationships for C. elegans genes. GenAge includes lifespan effects and cross references to NCBI Gene, as well as other databases. GenDR includes DR-induced Gene-phenotype associations.

5.3.2 High scoring genes regulate aging in *C. elegans*

Of the 48,231 *C. elegans* genes evaluated by HyQue for their role in aging, 1 gene received the highest score of 0.89 and 7 genes received a score of 0.78. Table 11 lists the genes with their WormBase identifier and gene symbol, as well as the HyQue data evaluation functions that contributed to their high evaluation score (where the function identifier corresponds to those in the list above). All of these genes have been reported in the literature to regulate longevity (PMID column of Table 11). The overall score distribution of all *C. elegans* genes is shown in Table 12.

Table 11 8 C. elegans genes that received the highest HyQue evaluations for their role in aging, the PubMed identifiers of papers describing their roles in regulating longevity, and the data evaluation functions that contributed to their scores

WormBase ID	Symbol	Score	PMIDs	Satisfied data evaluation function								
				1	2	3	4	5	6	7	8	9
WBGene00008205	sams-1	0.89	16103914	✓	✓	✓	✓	✓	✓		✓	✓
WBGene00009741	drr-1	0.78	16103914	✓	✓		✓	✓	✓		✓	✓
WBGene00000371	cco-1	0.78	21215371	✓	✓			✓	✓	✓	✓	✓
WBGene00002178	jnk-1	0.78	15767565	✓	✓			✓	✓	✓	✓	✓
WBGene00004800	sir-2.1	0.78	21938067	✓			✓	✓	✓	✓	✓	✓
WBGene00004789	sgk-1	0.78	15068796	✓	✓			✓	✓	✓	✓	✓
WBGene00006796	unc-62	0.78	17411345	✓	✓			✓	✓	✓	✓	✓
WBGene00004013	pha-4	0.78	19239417		✓		✓	✓	✓	✓	✓	✓

Table 12 HyQue score distribution for 48,231 C. elegans genes

Score	Number of genes
0.89	1
0.78	7
0.67	46
0.56	122
0.44	333
0.33	537
0.22	1759
0.11	9200
0	36226

HyQue captures the individual score contributions of each data evaluation function for each *C. elegans* gene, and using this data we measured the frequencies with which each function was satisfied across all *C. elegans* genes. Table 13 shows the frequency with which each of the 9 data evaluation functions were satisfied across all *C. elegans* genes.

Table 13 Frequency with which each data evaluation function was satisfied across all 48,231 C. elegans genes

Data evaluation function	**Satisfied frequency**	**Proportion** (Frequency/# of *C. elegans* genes)
DEF1	317	6.6×10^{-3}
DEF2	6406	1.3×10^{-1}
DEF3	1	2.1×10^{-5}
DEF4	55	1.1×10^{-3}
DEF5	699	1.4×10^{-2}
DEF6	876	1.8×10^{-2}
DEF7	135	2.8×10^{-3}
DEF8	1216	2.5×10^{-2}
DEF9	6899	1.4×10^{-1}

We also compared HyQue's evaluations of genes that would be expected to receive a high score to its evaluation of all genes, based on a naïve analysis of the gene descriptions in WormBase. Specifically, we queried the Bio2RDF WormBase dataset for genes that have at least one of the following terms in their WormBase description: "aging", "lifespan", "life span" and "longevity", which returned a set of 209 genes. The distribution of HyQue scores for this set of genes is significantly different from the distribution of HyQue scores for all *C. elegans* genes (Kolmogorov-Smirnov test $p<2.2 \times 10^{-16}$; see Figure 12). The score distribution of all *C. elegans* genes is heavily left skewed, with 0 being the most frequently assigned score. In contrast, of the 209 genes with aging-related terms in their description, the most frequently assigned score is 0.44

and >50% are assigned that score or higher (comparatively, just over 1% of all genes have a score of 0.44 or higher).

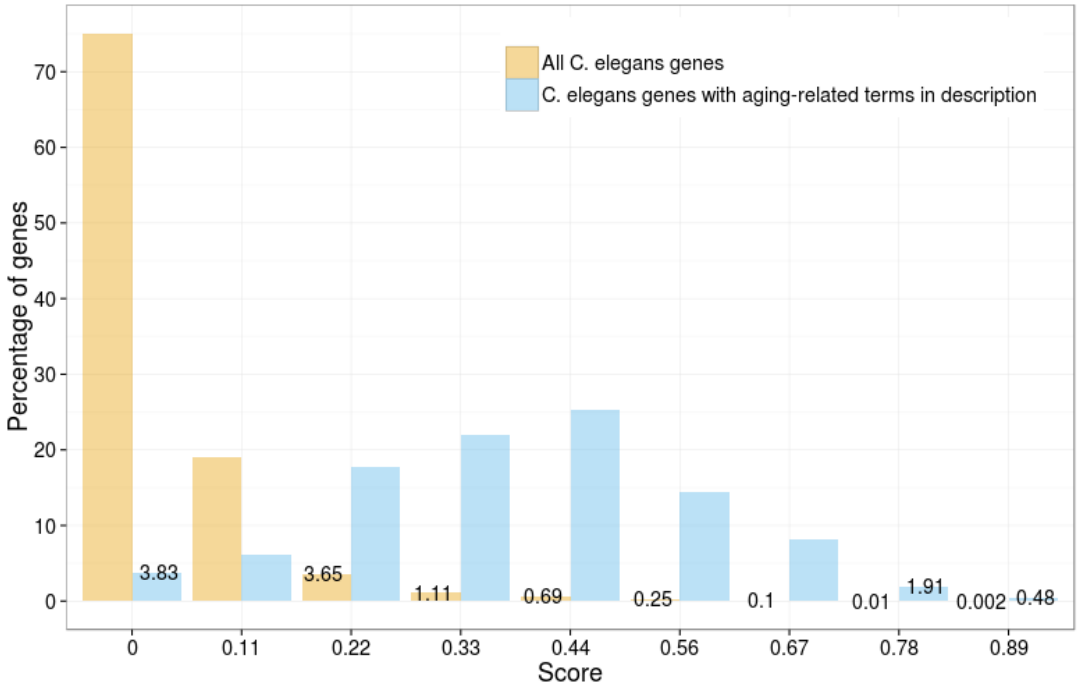


Figure 12 **The HyQue score distribution of all *C. elegans* genes is significantly different from that of the scores of 209 genes with aging-related terms in their WormBase descriptions (Kolmogorov-Smirnov test p<2.2x10-16). The percentage of genes with a given score is displayed when < 5%.**

5.3.3 HyQue identifies candidate aging-related genes in *C. elegans*

The analysis of HyQue scores for genes with aging-related terms in their descriptions gives us confidence that HyQue can discern aging-related genes from non-aging-related genes, and our next step was to use HyQue to identify genes that are good candidates for having a role in aging but which have not been characterized as such. To identify genes that are strong candidates for being aging/longevity-related we analyzed the HyQue evaluations to find the highest scoring genes that do not have an existing aging- or longevity-associated annotation in WormBase or GenAge (*i.e.* those genes that do not have a scoring contribution from DEF1 or DEF5), indicating that their involvement in

aging is not well characterized. Of the 509 genes with scores of 0.44 or greater, there are 31 such genes. Table 14 lists these genes with the data evaluation functions contributing to their overall score.

Table 14 31 C. elegans genes that received High HyQue evaluation scores for their role in aging without existing aging-related annotations, and the data evaluation functions that contributed to their scores.

WormBase ID	Symbol	Satisfied data evaluation function								
		1	2	3	4	5	6	7	8	9
WBGene00000252	bli-2		✓					✓	✓	✓
WBGene00000255	bli-5		✓				✓		✓	✓
WBGene00000262	bra-1		✓					✓	✓	✓
WBGene00000479	cgh-1						✓	✓	✓	✓
WBGene00000915*	daf-21						✓	✓	✓	✓
WBGene00001165	efn-4		✓					✓	✓	✓
WBGene00001428*	fkb-3		✓				✓		✓	✓
WBGene00001543*	gcy-18		✓				✓		✓	✓
WBGene00001578	ges-1		✓				✓		✓	✓
WBGene00001746	gsk-3		✓					✓	✓	✓
WBGene00001824	hbl-1		✓				✓		✓	✓
WBGene00001974	hmg-4		✓					✓	✓	✓
WBGene00001979	hmp-2		✓					✓	✓	✓
WBGene00002005*	hsp-1				✓		✓		✓	✓
WBGene00002013*	hsp-12.6		✓		✓				✓	✓
WBGene00002069*	ikb-1		✓		✓				✓	✓
WBGene00002881	let-756		✓					✓	✓	✓
WBGene00003029	lin-44		✓					✓	✓	✓
WBGene00003058	lov-1		✓					✓	✓	✓
WBGene00003210	mel-28						✓	✓	✓	✓
WBGene00003473	mtl-1		✓				✓		✓	✓
WBGene00003497	mup-4		✓					✓	✓	✓
WBGene00003977*	pes-2.1		✓				✓		✓	✓
WBGene00004392	rnr-2						✓	✓	✓	✓
WBGene00004765	sel-8		✓					✓	✓	✓
WBGene00006789	unc-54		✓					✓	✓	✓
WBGene00007036	sod-5		✓				✓		✓	✓
WBGene00016140	rpb-2						✓	✓	✓	✓
WBGene00017830	rpb-8		✓					✓	✓	✓
WBGene00020100	mks-1		✓				✓		✓	✓
WBGene00021334	vps-4		✓					✓	✓	✓

*** true positive**

A closer examination of these 31 genes revealed that there are 7 'true positive' cases (marked with a * in Table 14). Specifically, these 7 genes have WormBase human-readable descriptions that directly implicate them as aging/lifespan associated genes and they are annotated with the WormBase 'lifespan variant' phenotype. daf-21 encodes an Hsp90-family molecular chaperone known to regulate dauer formation [144], and its RNAi-induced under expression reduces age-1 modulated lifespan [145]. fkb-3 encodes a peptidylprolyl cis/trans isomerase and its expression is positively regulated by the DAF-2 pathway and DAF-16 FOXO transcription factor activity [146]. gcy-18 encodes a guanyl cyclase crucial for wild-type thermotaxis [147]. gcy-18 expression is induced in DAF-2/DAF-16 double mutants, and its knockout by RNAi extends lifespan [148]. hsp-1 encodes heat shock protein hsp70A and its knockout by RNAi reduces lifespan in an age-1 mutant [144]. hsp-12.6 is a stress response gene downstream of daf-16. hsp-12.6 expression is increased in daf-2 mutants [124, 148] and its silencing by RNAi reduces lifespan by approximately 25% [122]. ikb-1 deletion mutants also have a shortened lifespan [149] and ikb-1 function may be related to DNA damage response [150]. pes-2.1 expression is down-regulated in daf-2 loss-of-function mutants, and RNAi targeting of pes-2.1 increases *C. elegans* lifespan [148]. 3 other genes – ges-1, mtl-1 and sod-5 – have the 'lifespan variant' phenotype but their roles in aging/longevity are not characterized.

The most frequently occurring combination of satisfied data evaluation functions (used to generate the overall HyQue evaluation score) for the 31 candidate genes is DEF2, DEF7, DEF8 and DEF9. Based on the overall frequencies of these functions being satisfied (Table 12) across all *C. elegans* genes, the likelihood of observing this combination by chance for a single gene is just 1.34×10^{-6}. Considering that there are

48,231 genes in WormBase, less than one gene in this set would have this combination by chance.

We used FUNC [151] to analyze the 31 candidate genes for significantly enriched biological function and process annotations from GOA. Enriched biological process terms include 'determination of adult lifespan' (GO:0008340; p-value* 2.7×10^{-11}), 'nematode larval development' (GO:0002119, p-value 4.6×10^{-3}), 'multicellular organismal protein metabolic process' (GO:0044268, p value 5.7×10^{-3}), 'response to heat' (GO:0009408, p value 6.7×10^{-3}), 'larval foraging behavior' (GO:0035177, p value 1.1×10^{-2}), 'multicellular organismal reproductive process' (GO:0048609, p value 1.3×10^{-2}), 'superoxide metabolic process' (GO:0006801, p value 1.7×10^{-2}), 'inositol lipid-mediated signaling' (GO:0048017, p value 1.9×10^{-2}), 'deoxyribonucleoside diphosphate metabolic process' (GO:0009186, p value 2.5×10^{-2}) and 'protein folding' (GO:0006457, p value 2.6×10^{-2}). Enriched molecular function terms include 'structural constituent of collagen and cuticulin-based cuticle' (GO:0042329, p value 2.9×10^{-4}), 'fibroblast growth factor receptor binding' (GO:00005104 p value 5.2×10^{-3}), 'superoxide dismutase activity' (GO:0004784, p value 1.2×10^{-2}), 'growth factor activity' (GO:0008083, p value 2.1×10^{-2}), and 'transcription coactivator activity' (GO:0003713, p value 3.6×10^{-2}). Supplementary Table 16 and Table 17 in Appendix B list all significantly enriched GO biological process and molecular function annotations and their associated p-values.

5.4 Discussion

We have demonstrated that HyQue is able to correctly identify known aging/lifespan-related genes in *C. elegans* by evaluating a variety of evidence types from multiple

* p-values were calculated after a FUNC refinement step to remove GO terms that were enriched only because their child terms were enriched

sources, and can also identify candidate aging and longevity-related genes whose effect on these biological processes are not yet well-characterized. Indeed, the 24 candidate genes (not including the 7 true positive candidates) are promising targets for future research to uncover their effects on lifespan. The HyQue data evaluation functions that were not satisfied for each of these genes can be used as a guide for future experimental designs (for example, given that experimental data about the expression of gsk-3 under dietary restriction is not currently available, an experiment could be performed to obtain this data). Of the 24 candidate genes lacking direct links to aging/lifespan, the majority have known functions related to development, stress response (including protection against environmental stresses such as heat and oxidative damage) and reproductive behavior in *C. elegans*. Human orthologs [152] of several of these genes are also responsible for neurodegenerative disease phenotypes. For example, polymorphisms in a human ortholog of let-756, FGF20, are risk factors for Parkinson's disease [153, 154]. Similarly, a human ortholog of gsk-3, GSK3B, may also modulate risk for Parkinson's disease [155] and Alzheimer's disease [156, 157]. Mutations in SOD1, a human ortholog of sod-5 that functions to destroy free superoxide radicals in the body and protect against RNA, DNA and protein damage, are associated with amyotrophic lateral sclerosis (ALS, or Lou Gehrig's disease). All of these human disorders are associated with shortened lifespan [158-160] .

No *C. elegans* genes achieved the maximum possible normalized HyQue score of 1, because no single gene had all features required to satisfy the 9 data evaluation functions used in its evaluation. More specifically, only one *C. elegans* gene satisfied DEF3, which asked if a given gene had a homolog significantly differently expressed

under dietary restriction across multiple studies, using the Bio2RDF GenDR dataset as a source. GenDR includes the homologs of 99 model organism genes, but only three of these had entries in the GenDR list of DR affected genes from multiple studies, and of those only one was a homolog of a *C. elegans* gene. This dataset will be important for future applications of HyQue, however, as we extend its application to evaluating the role of mammalian genes in aging in a manner similar to the approach described here. Also, if HAGR extends the GenDR database to non-mammals, then it will also gain relevance for HyQue.

The frequency with which each of the HyQue data evaluation functions that contributed to the scores of the 31 candidate genes were satisfied in all *C. elegans* genes indicates that is very unlikely for a gene to have satisfied this set of functions by chance. This, as well as the distribution of HyQue scores for genes with aging-related terms in their descriptions that were expected to receive an evaluation reflecting their role in aging-related biological processes validates the HyQue approach to assessing biological hypotheses, whereby genes that have accumulated more biological evidence (as determined by the execution of data retrieval and evaluation functions) are better candidates for satisfying the hypothesis that they are involved in aging and thus receive a higher evaluation score. The set of 209 genes with aging-related terms in their description do not comprise *all* genes that have a role in aging (for example, it includes only 5 of the 8 top-scoring genes – cco-1, jnk-1, sgk-1, sams-1 and pha-4), but the occurrence of aging-related terms in these genes' descriptions implicates them as aging-related genes, and thus HyQue was expected to evaluate them as such and assign them higher evaluation scores than would be expected by chance, as was observed.

The scoring system used by HyQue to evaluate a gene's role in aging is one of many possible variations, and will improve over time. For example, currently all evidence types are assigned the same weight, and so the presence or absence of any evidence equally affects HyQue's final evaluation. However, some evidence, such as experimentally measured changes in gene expression, may have more validity in confirming or refuting a hypothesis. This could be reflected by, for example, increasing the score contributed by gene expression data so that its value affects a final score more than a less powerful data source, such as a one-step-removed genetic interaction. It may also be that different scientists will come to view the same evidence with varying confidence, and HyQue's evaluation functions can evolve over time to reflect these shifts in perspective. HyQue's automatically generated provenance of hypothesis evaluations is useful in this context, as it makes it possible to determine exactly how a hypothesis achieved a given score, by following links to evaluation rules and source data. Data retrieval and data evaluation are separated to facilitate the re-use of data retrieval functions for different hypothesis types, and also in an attempt to "future proof" HyQue functions in the event that a data source changes, or a data evaluation criteria changes over time. Maintaining data retrieval and evaluation functions separately means that either can be updated without requiring that the other be changed.

Performance evaluation measures such as veracity [161], recently proposed as an alternative approach to precision and recall for evaluating predictive systems, may also be useful in assessing HyQue's ability to correctly evaluate hypotheses. Veracity quantifies the performance of systems that predict features such as a chemical's toxicological activity by considering what proportion of a set of entities that are input into

the system should ideally fall into each of the possible predicted categories, and comparing the actual predicted proportions to this ideal. In other words, veracity quantifies the confidence level associated with a given prediction, in that we can have more confidence in predictions that more closely follow the ideal distribution. Using veracity to assess HyQue's evaluations of *C. elegans* genes for their involvement in aging would require an ideal distribution of scores, which would in part require verification of each gene's role (or lack thereof) in aging. Such an assessment may be possible in the near future.

HyQue realizes the promise of the Semantic Web [10] to bring relevant knowledge automatically to the fingers of biologists studying complex domains, and to reason over this knowledge for assessing biological hypotheses. With current biological evidence, the functions used to evaluate that evidence, and the outcomes of HyQue evaluations all serialized as Linked Data, it is possible to query and reason over these resources to discover how evidence changes over time, and how this affects prevailing biological hypotheses. HyQue data retrieval and evaluation SPIN functions can also be repurposed for new biological domains, and their availability as linked data whose properties can be computationally queried (for example, to discover functions that satisfy a given criteria or retrieve a certain data type) makes them ideal for re-use.

Future work will involve experimental validation of the 24 candidate genes for their role in lifespan-related biological processes, as well as continued development of the data retrieval and evaluation functions used by HyQue to assess a gene's role in aging. Specifically, it is possible to expand the taxonomic reach of HyQue by including evidence from additional model organisms as such data becomes available.

In summary, we have described the application of HyQue, a Semantic Web tool for hypothesis evaluation, to the problem of discovering genes that affect aging and longevity. We show that HyQue gives positive scores to hypotheses involving genes that are known to regulate aging, and also identified 24 potentially aging-related genes that are good candidates for experimental study in this context. The rules and functions we developed for this domain can be re-used in future applications of HyQue.

6 Chapter: Summary of contributions and future directions

The research presented in this thesis lays a foundation for and describes preliminary results on the use of a Semantic Web approach to computational hypothesis evaluation in the life sciences. Such an approach requires the coordination of tools for querying structured life sciences data, ontologies for reasoning about that data, rules for encoding potentially complex evaluation processes that consume these resources, and a model for recording the provenance of such tasks, into an ecosystem that has Semantic Web compatible data as both its input and output. Early attempts at such systems suffered from a lack of standards and the ability to publish to the Web in a machine-understandable way, making them inherently difficult to share and repurpose. This is changing rapidly with recent advances in infrastructure and applications for the Internet and Web [162], the exponential growth of published (but largely unstructured) biological data [3, 163], as well as the growing momentum of real-world Semantic Web applications in the sciences [18, 164, 165]. All of these factors have made the time ripe for designing and assessing a Semantic Web framework for hypothesis evaluation, focused on the life sciences. From this perspective, the work described in this thesis can be summarized into three main contributions:

- *Novel framing of the hypothesis evaluation task* as a combination of targeted scalable data retrieval coupled with automated analysis of that data in the context of domain knowledge
- *Design and implementation of HyQue, a Semantic Web tool for hypothesis evaluation* – in HyQue, data and domain knowledge are encoded as RDF linked data and OWL ontologies, hypothesis evaluation rules are described using the

SPARQL Inferencing Notation and encoded as RDF, and hypothesis evaluations and their provenance are encoded as RDF linked data

- *Application of HyQue to discover candidate C. elegans genes involved in aging* – using HyQue with data retrieval and evaluation functions developed for the domain of aging, we evaluated all 48,231 *C. elegans* genes and identified 24 candidate genes that had significantly positive HyQue evaluation scores but were not well characterized in terms of their effects on lifespan

In addition to these core contributions, my doctoral research has also contributed significant development of the Bio2RDF project, as well as the ovopub, a new model for data provenance on the Semantic Web with specific relevance to life sciences data. In the following pages, I revisit each of these contributions in more detail and propose future work.

In work on the Bio2RDF project, we took major strides towards making life sciences linked data available at a large scale. In particular, we used a dataset registry to coerce the naming of data items so that they connect together. We also demonstrated ontology-based integration through mappings between Bio2RDF dataset-specific ontologies and the Semanticscience Integrated Ontology (SIO). This enabled us to query and reason about Bio2RDF data using a common terminology. The migration of Bio2RDF open-source scripts to GitHub lowers the barrier for community participation in maintaining up-to-date releases of Bio2RDF datasets and contributing new datasets. Lastly, we demonstrate the value of Bio2RDF in answering questions and evaluating biological hypotheses.

In developing the ovopub for recording the provenance of statements and datasets we addressed a question of growing interest to the Semantic Web community – can we design a scalable, domain independent model for recording provenance on the Semantic Web? Having a machine-understandable description of data provenance is increasingly important as the number of tools and projects that consume data on the Semantic Web continues to grow, and being able to discriminate data of quality depends on being able to assess its source, how it was made, and its currency. The ovopub can be used to describe the provenance of statements and datasets of any size or level of complexity, and also enables provenance-based querying and selection. HyQue uses the ovopub model to record the provenance of each hypothesis evaluation it performs, which will facilitate provenance-based assessment of HyQue performance.

HyQue has seen many iterations, the most recent resulting in the system described in the methods sections of Chapters 2 and 5. What started out as a system whose logic and rules were encoded almost entirely in locally executed PHP scripts has become a full-fledged Semantic Web tool that both consumes and produces RDF linked data, and is itself largely described using RDF through the SPIN modeling vocabulary. In parallel, our early applications of HyQue to the galactose metabolism gene network in *S. cerevisiae*, inspired by the work of HyBrow, have evolved to a far more advanced application toward evaluating hypotheses about genes and their role in aging and longevity in the model organism *C. elegans*. Using HyQue, we both re-confirmed experimental and annotation-based support for genes known to modulate aging and longevity, and discovered 24 candidate genes that are promising subjects for further experimentation.

Future work with HyQue lies in several areas: experimentally validating HyQue evaluations, extending the application domains of HyQue, and improving HyQue's user interface and user experience. The biggest 'missing piece' of the HyQue story lies in experimentally assessing HyQue evaluations, specifically the 24 candidate aging-related genes identified by HyQue as described in Chapter 5. Statistical analysis of HyQue's evaluation of all *C. elegans* genes demonstrates that the system is capable of distinguishing and accurately evaluating aging-related genes in comparison to the population of all genes, and so experimental validation is the next logical step. Experimental approaches for measuring changes in gene expression and lifespan under different experimental conditions in *C. elegans* have been well described (*e.g.* [166-169]), making the design and execution of such experiments focused on the 24 candidate genes a feasible next step, in collaboration with domain experts.

We have also completed preliminary work in extending the application of HyQue to the domain of drug safety, and more specifically drug-induced cardiotoxicity, the leading reason for drug recalls from 2002 to 2011 [170]. We are collaborating with researchers at the U.S. Food and Drug Administration to use HyQue to evaluate hypotheses about the cardiotoxic effects of tyrosine kinase inhibitors (TKIs) [171] which are used in cancer therapies to inhibit tumour growth. Different TKIs have different cardiotoxic effects, and the goal of using HyQue is to discern which TKI features may be associated with cardiotoxicity by taking advantage of the large and disparate sources of relevant data available in a scalable manner, addressing a need of the FDA researchers who have abundance of data but lack methods for querying, integrating and reasoning over it. We have developed domain specific rules for HyQue that assess evidence from

the ChEMBL database of bio-assays, drug information sources including the Comparative Toxicogenomics Database, DrugBank, and SIDER about drug side effects and drug-drug interactions, as well as results from mouse model experiments and human clinical trials. We are currently assessing the application of these rules to evaluate tyrosine kinase inhibitors that have known cardiotoxic effects (as a proof-of-concept), as well as evaluate hypotheses about the cardiotoxic effects of non-TKI drugs.

Another area for future work lies in developing an online graphical user interface for HyQue that allows users to submit hypotheses, view and select evaluation rules, and also view evaluation results. I have developed a Drupal module for HyQue that allows users to compose and submit hypotheses for evaluation (Figure 13), and also a preliminary interface for visualizing drug-cardiotoxicity hypothesis evaluation results (Figure 14). Extensions to this early work will involve rendering the evaluation results display interface more flexible to facilitate its re-use for displaying HyQue hypothesis evaluation results for any domain or hypothesis type.

In addition to the proposed future work specific to the implementation of HyQue and assessing its evaluations, there are also more exploratory research questions motivated by the HyQue approach to hypothesis evaluation. For example, it remains an open question how the rules, data evaluation functions and data retrieval functions that HyQue uses, which are currently curated and choreographed by human scientists, may be reasoned over, automatically selected and combined on the fly by Semantic Web software agents, based on input hypothesis features. The Semantic Automated Discovery

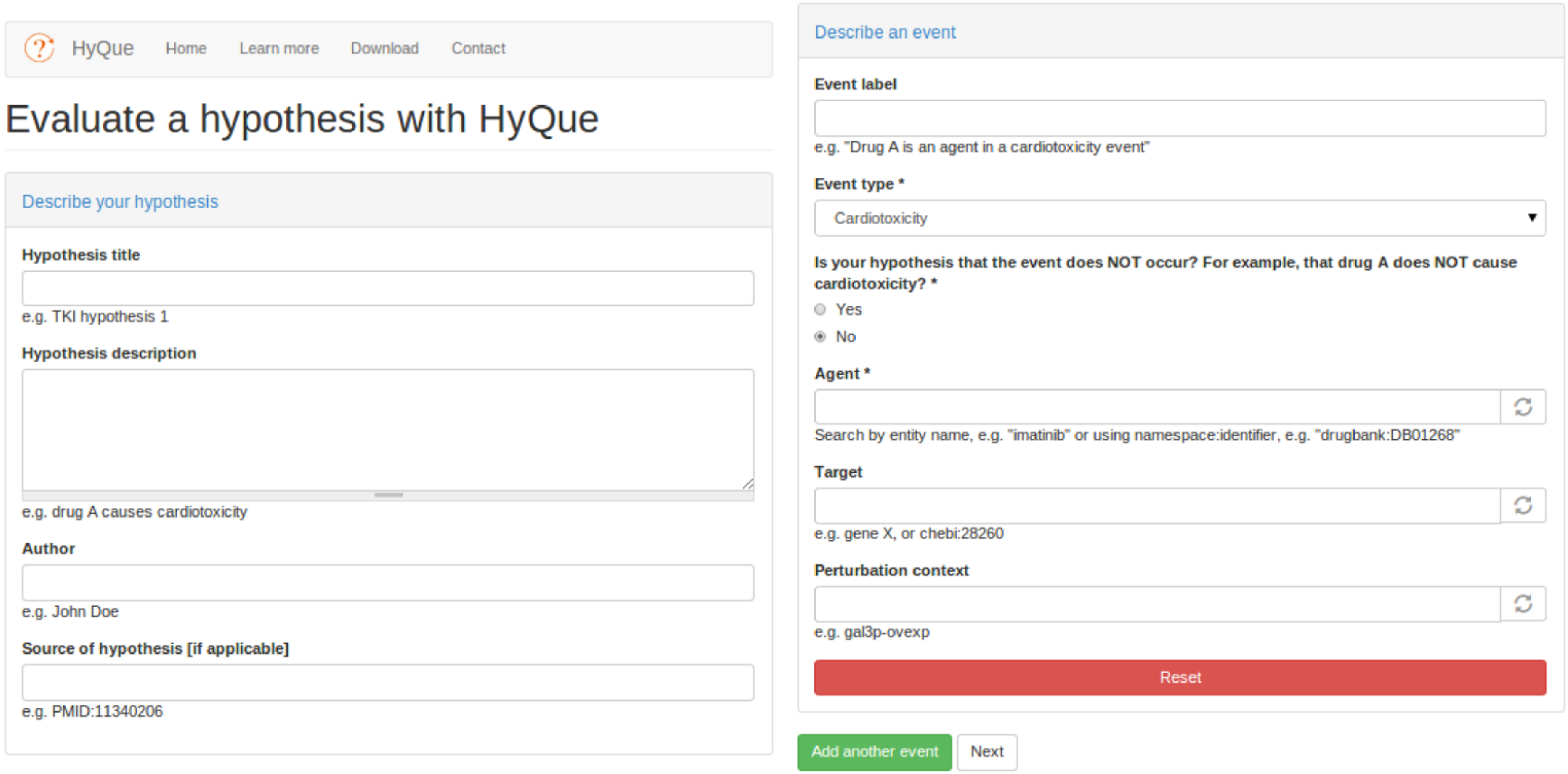


Figure 13 Form-based HyQue user interface for composing hypotheses about drug cardiotoxicity, implemented as a Drupal module.

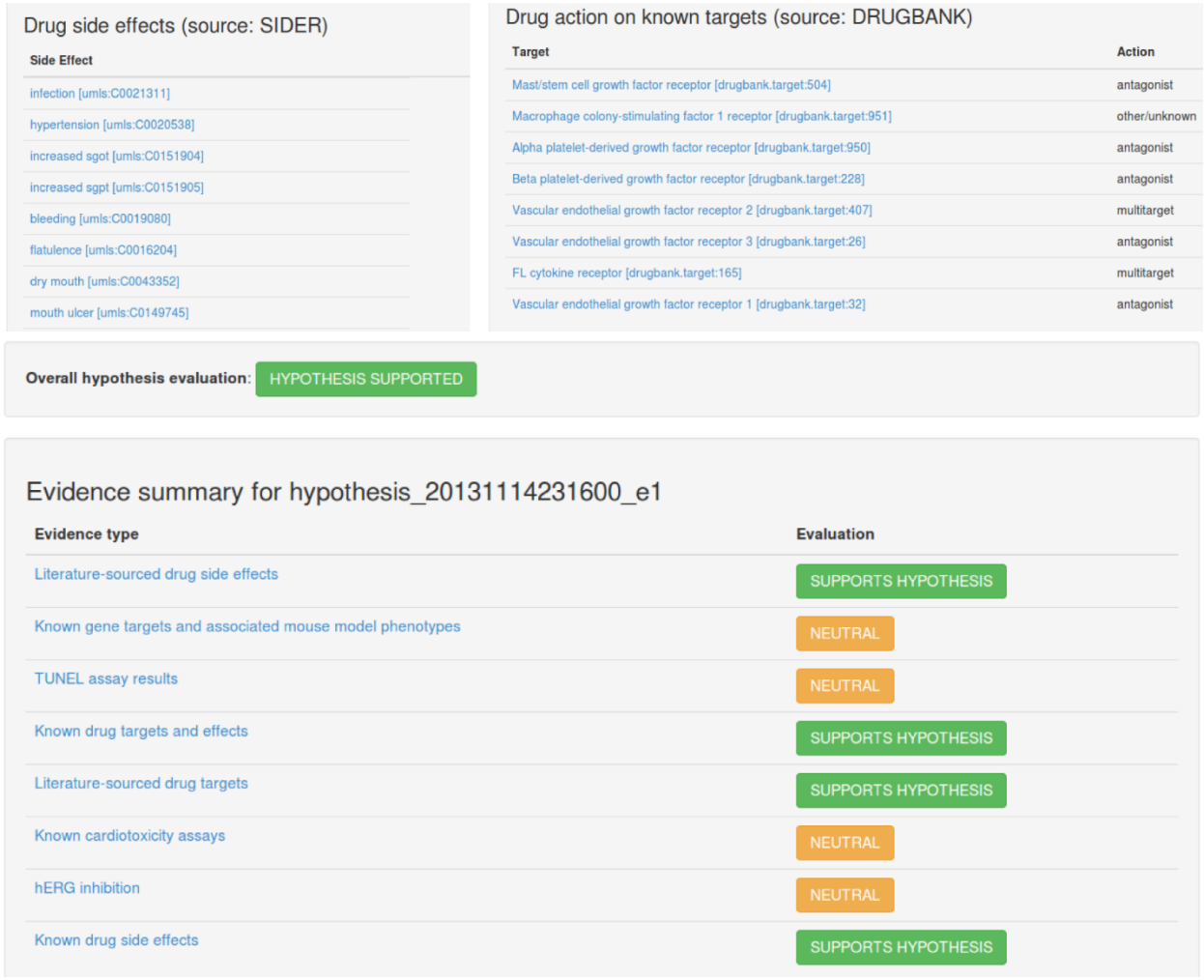


Figure 14 HyQue user interface for displaying drug cardiotoxicity hypothesis evaluation results, including data retrieved and contribution of different evidence types to overall evaluation.

and Integration (SADI) framework [172] is a promising candidate for exploring this potential application and extension of HyQue. The SADI framework allows users to formally describe the inputs and outputs of web services using OWL class definitions, and generates code stubs for required methods based on OWL SADI service descriptions. SADI has been used to develop web services for classifying small molecules [173, 174] and for a text mining pipeline [175] that extracts genetic mutants and their phenotypes from scientific publications. Future research could explore approaches for automatically selecting data retrieval and evaluation functions through SADI web services triggered by hypotheses that instantiate the HyQue Hypothesis Ontology. As a basis for this functionality, we are interested in also developing databases for storing and publishing HyQue hypotheses, data retrieval functions, and data evaluation functions, as well as their provenance.

The motivating use cases for HyQue consider the task of hypothesis evaluation as one of retrieving data identified by experts to be relevant for the hypothesis at hand, and assessing that data using evaluation functions that are domain specific and potentially unique to that same expert. The data evaluation functions are then combined to generate scores where the score is linearly related to the amount of supporting evidence. In contrast, machine learning approaches for predicting the association of a biological entity with a property of interest (*e.g.*, whether or not a gene affects lifespan) use labeled gold standard training data to statistically determine which features and feature values are critical to correctly predicting the property, and weight them accordingly [176]. In the domain of biogerontology the application area for HyQue described in Chapter 5, labeled data that could be used as a gold standard is be the set of genes with aging-related terms

in their human-readable descriptions. HyQue could be integrated with a machine learning approach to prediction by acting as a tool for feature vector creation via data retrieval specified by an input hypothesis, the predictive value of which could then be determined by a given machine learning algorithm. More specifically, the features identified by a classifier as significant could be assigned a higher scoring weight by HyQue when calculating hypothesis scores. This has potential of avoiding non-optimal (*i.e.* less predictive) scoring functions while taking advantage of HyQue's approach to targeted data retrieval that is semantically linked with specific hypothesis types.

The ultimate goal of this research has been to assist biologists in evaluating hypotheses in the data-rich setting that is science today. With HyQue, we achieve this by making targeted data retrieval scalable in order to leverage the 'unreasonable effectiveness' of large datasets [177], and by automating the analysis of retrieved data in the context of domain knowledge. The rules and functions used by HyQue have been developed through collaboration between domain experts of two kinds – subject matter experts familiar with the field of aging research, and bioinformatics and Semantic Web experts knowledgeable in the areas of biological data integration, querying and reasoning using Semantic Web tools. This approach to rule development for HyQue contrasts computational approaches for automatically learning rules from data that rely on statistical analysis and require little to no human input. The advantage of the latter is that automatic approaches can efficiently operate over very large amounts of data and produce rules that perform well in predicting features of interest. However, such automatic approaches also face significant challenges: the quality of rules produced depends on the quality of the source data, and on the ability to correctly identify relationships between

data items. Biological data generated by different scientists, and published in a variety of formats are not easily integrated without work by bioinformaticians to manage their idiosyncrasies and translate the semantics of the data to a machine understandable format, as described in Chapter 3. Thus, the ability to process large amounts of data does not guarantee the success of rule learning approaches. From the user perspective, methods to learn rules from data can generate rules that may have significant positive predictive value, but are completely opaque to those interpreting the rules and attempting to improve their understanding of a biological system using their predictions. In contrast, because HyQue rules are manually crafted, can be re-used, and because the HyQue framework captures exactly how rules are used to evaluate a given hypothesis, the reasoning processes employed by HyQue through its rules are highly transparent and interpretable by its users. Challenges remain in effectively sharing and making discoverable the components of HyQue that have potential for re-use in other biological domains, and in continuing to lower the barrier of effective use of HyQue, and Semantic Web technologies in general, by the experimental biologist. The target user for the current version of HyQue is the bioinformatician, who though not necessarily an expert in Semantic Web technologies, is versed in developing and applying computational approaches for data management and analysis often in collaboration with biologists who have a specific use case or question. Making HyQue more directly accessible to the biologist will require development of a more flexible user interface for crafting and evaluating hypotheses, a browse and search tool for identifying and selecting rules relevant to a given user, as well as a tool for crafting functions and rules that relies less on experience with SPARQL and SPIN. The promise of the Semantic Web for real-world

applications in the life sciences has never been greater: a future version of HyQue improved by biologist feedback and application to broader domains, and, more importantly, the biological discoveries that HyQue can facilitate, are exciting prospects.

Appendices

Appendix A

Table 15 Protein-protein interaction detection methods used by DEF7 to filter results

Bio2RDF URI	Name
http://bio2rdf.org/psi-mi:0004	affinity chromatography technology
http://bio2rdf.org/psi-mi:0006	anti bait coimmunoprecipitation
http://bio2rdf.org/psi-mi:0007	anti tag coimmunoprecipitation
http://bio2rdf.org/psi-mi:0012	bioluminescence resonance energy transfer
http://bio2rdf.org/psi-mi:0019	coimmunoprecipitation
http://bio2rdf.org/psi-mi:0020	transmission electron microscopy
http://bio2rdf.org/psi-mi:0040	electron microscopy
http://bio2rdf.org/psi-mi:0055	fluorescent resonance energy transfer
http://bio2rdf.org/psi-mi:0067	light scattering
http://bio2rdf.org/psi-mi:0069	mass spectrometry studies of complexes
http://bio2rdf.org/psi-mi:0077	nuclear magnetic resonance
http://bio2rdf.org/psi-mi:0096	pull down affinity chromatography
http://bio2rdf.org/psi-mi:0107	surface plasmon resonance
http://bio2rdf.org/psi-mi:0109	tap tag coimmunoprecipitation
http://bio2rdf.org/psi-mi:0114	X-ray crystallography
http://bio2rdf.org/psi-mi:0254	genetic interference
http://bio2rdf.org/psi-mi:0364	inferred by curator
http://bio2rdf.org/psi-mi:0405	competition binding
http://bio2rdf.org/psi-mi:0406	deacetylase assay
http://bio2rdf.org/psi-mi:0410	electron tomography
http://bio2rdf.org/psi-mi:0411	enzyme linked immunosorbent assay
http://bio2rdf.org/psi-mi:0415	enzymatic study
http://bio2rdf.org/psi-mi:0417	footprinting
http://bio2rdf.org/psi-mi:0423	in-gel kinase assay
http://bio2rdf.org/psi-mi:0424	protein kinase assay
http://bio2rdf.org/psi-mi:0434	phosphatase assay
http://bio2rdf.org/psi-mi:0435	protease assay
http://bio2rdf.org/psi-mi:0515	methyltransferase assay
http://bio2rdf.org/psi-mi:0676	tandem affinity purification
http://bio2rdf.org/psi-mi:0678	antibody array
http://bio2rdf.org/psi-mi:0728	gal4 vp16 complementation
http://bio2rdf.org/psi-mi:0809	bimolecular fluorescence complementation
http://bio2rdf.org/psi-mi:0826	X-ray scattering
http://bio2rdf.org/psi-mi:0841	phosphotransfer assay
http://bio2rdf.org/psi-mi:0858	immunodepleted coimmunoprecipitation
http://bio2rdf.org/psi-mi:0870	demethylase assay

Appendix B

Table 16 GO biological process annotations enriched in the set of 31 *C. elegans* candidate aging-related genes identified by HyQue

Biological process	GO identifier	P-value
negative regulation of translation	GO:0017148	0.042
morphogenesis of embryonic epithelium	GO:0016331	0.034
multi-organism reproductive behavior	GO:0044705	0.034
engulfment of apoptotic cell	GO:0043652	0.030
protein folding	GO:0006457	0.026
mitotic spindle organization	GO:0007052	0.026
deoxyribonucleoside diphosphate metabolic process	GO:0009186	0.026
thermosensory behavior	GO:0040040	0.026
response to cadmium ion	GO:0046686	0.026
Wnt receptor signaling pathway, regulating spindle positioning	GO:0060069	0.026
sexual reproduction	GO:0019953	0.023
cuticle development involved in collagen and cuticulin-based cuticle molting cycle	GO:0042338	0.023
cytoskeletal anchoring at plasma membrane	GO:0007016	0.020
regulation of cell adhesion	GO:0030155	0.020
inositol lipid-mediated signaling	GO:0048017	0.020
superoxide metabolic process	GO:0006801	0.017
fibroblast growth factor receptor signaling pathway	GO:0008543	0.017
tail tip morphogenesis	GO:0045138	0.016
skeletal muscle myosin thick filament assembly	GO:0030241	0.014
negative regulation of transforming growth factor beta receptor signaling pathway	GO:0030512	0.014
regulation of axon extension involved in axon guidance	GO:0048841	0.014
negative regulation of synapse assembly	GO:0051964	0.014
gonad development	GO:0008406	0.014
multicellular organismal reproductive process	GO:0048609	0.013
negative regulation of cell projection organization	GO:0031345	0.011
larval foraging behavior	GO:0035177	0.011
retrograde transport, endosome to Golgi	GO:0042147	0.011
cell fate specification involved in pattern specification	GO:0060573	0.011

Biological process	GO identifier	P-value
receptor guanylyl cyclase signaling pathway	GO:0007168	0.0086
multicellular organismal protein catabolic process	GO:0044254	0.0086
Hatching	GO:0035188	0.0073
hermaphrodite genitalia development	GO:0040035	0.0072
response to heat	GO:0009408	0.0067
germline cell cycle switching, mitotic to meiotic cell cycle	GO:0051729	0.0064
multicellular organismal protein metabolic process	GO:0044268	0.0057
nematode larval development	GO:0002119	0.0046
regulation of actin cytoskeleton organization by cell-cell adhesion	GO:0090138	0.0029
apoptotic process	GO:0006915	0.0019
receptor-mediated endocytosis	GO:0006898	0.00094
morphogenesis of an epithelium	GO:0002009	0.00094
cell migration involved in gastrulation	GO:0042074	0.00093
Locomotion	GO:0040011	0.00030
ATP catabolic process	GO:0006200	0.00028
determination of adult lifespan	GO:0008340	2.17E-11

Table 17 GO molecular function annotations enriched in the set of 31 *C. elegans* candidate aging-related genes identified by HyQue

Molecular function	GO identifier	P-value
four-way junction helicase activity	GO:0009378	0.0481223
actin filament binding	GO:0051015	0.0406744
transcription coactivator activity	GO:0003713	0.0356782
ATP binding	GO:0005524	0.0240303
ATPase activity	GO:0016887	0.0212704
ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor	GO:0004748	0.0205395
growth factor activity	GO:0008083	0.0205395
frizzled binding	GO:0005109	0.0154429
RNA helicase activity	GO:0003724	0.012885
superoxide dismutase activity	GO:0004784	0.012885
structural constituent of muscle	GO:0008307	0.012885
cadherin binding	GO:0045296	0.0103208
unfolded protein binding	GO:0051082	0.0084738
protein kinase binding	GO:0019901	0.0078768
microfilament motor activity	GO:0000146	0.00517324

Molecular function	GO identifier	P-value
fibroblast growth factor receptor binding	GO:0005104	0.00517324
protein domain specific binding	GO:0019904	0.00405088
alpha-catenin binding	GO:0045294	0.00258983
structural constituent of collagen and cuticulin-based cuticle	GO:0042329	0.000285639

References

1. Brandon RN: **Does biology have laws?** *Philosophy of Science* 1997, **64**:S444-S457.
2. Neylon C, Wu S: **Article-Level Metrics and the Evolution of Scientific Impact**. *PloS Biology* 2009, **7**(11):e1000242.
3. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S *et al*: **Big data: The future of biocuration**. *Nature* 2008, **455**(7209):47-50.
4. The Royal Society: **Science as an open enterprise** [http://royalsociety.org/policy/projects/science-public-enterprise/report/].
5. Bourne PE: **What Big Data means to me**. *Journal of the American Medical Informatics Association : JAMIA* 2014, **21**(2):194.
6. Goble C: **The low down on e-science and grids for biology**. *Comp Funct Genomics* 2001, **2**(6):365-370.
7. Rauwerda H, Roos M, Hertzberger BO, Breit TM: **The promise of a virtual lab in drug discovery**. *Drug Discov Today* 2006, **11**(5-6):228-236.
8. Narock T, Fox P: **From science to e-Science to Semantic e-Science: A Heliophysics case study**. *Computers & Geosciences* 2012, **46**:248-254.
9. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, Couch P, Cruickshank D, Delderfield M, Dunlop I *et al*: **Why linked data is not enough for scientists**. *Future Generation Computer Systems* 2013, **29**(2):599-611.
10. Berners-Lee T, Hendler J, Lassila O: **The Semantic Web**. *Scientific American* 2001, **284**(5):34-43.

11. Bratt S: **Semantic Web Stack** [http://www.w3.org/2004/Talks/1117-sb-gartnerWS/slide18-0.html].

12. Heath T, Bizer C: **Linked Data: Evolving the Web into a Global Data Space**. *Synthesis Lectures on the Semantic Web: Theory and Technology* 2011, **1**(1):1-136.

13. Slater T, Bouton C, Huang ES: **Beyond data integration**. *Drug discovery today* 2008, **13**(13-14):584-589.

14. Antezana E, Kuiper M, Mironov V: **Biological knowledge management: the emerging role of the Semantic Web technologies**. *Brief Bioinform* 2009, **10**(4):392-407.

15. Shah N, Musen MA: **Ontologies in support of formal representations of biological systems**. In: *The Handbook on Ontologies (2nd edition)*. Edited by Staab S, Studer R: Springer Berlin Heidelberg; 2010: 445-461.

16. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems**. *Journal of Biomedical Informatics* 2008, **41**(5):706-716.

17. Nolin M-A, Ansell P, Belleau F, Idehen K, Rigault P, Tourigny N, Roe P, Hogan JM, Dumontier M: **Bio2RDF Network of Linked Data**. In: *Semantic Web Challenge; International Semantic Web Conference (ISWC 2008); Karlsruhe, Germany*. 2008.

18. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C *et al*: **Open PHACTS: semantic**

interoperability for drug discovery. *Drug Discov Today* 2012, **17**(21-22):1188-1198.

19. Gruber T: **Toward principles for the design of ontologies used for knowledge sharing**. *International Journal of Human-Computer Studies* 1995, **43**(5-6):907-928.

20. Cuenca Grau B, Horrocks I, Motik B, Parsia B, Patel-Scheider PF, Sattler U: **OWL 2: The next step for OWL**. *Web Semantics: Science, Services and Agents on the World Wide Web* 2008, **6**:309-322.

21. W3C: **OWL 2 Web Ontology Language Primer (Second Edition)** [http://www.w3.org/TR/owl2-primer].

22. Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Scheider PF (eds.): **The Description Logic Handbook: Theory, Implementation and Applications**: Cambridge University Press; 2007.

23. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

24. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology**. *Nucleic Acids Res* 2004, **32**(Database issue):D262-266.

25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **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* 2005, **102**(43):15545-15550.

26. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A *et al*: **A large-scale evaluation of computational protein function prediction**. *Nature methods* 2013, **10**(3):221-227.

27. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, Ferris TA, Shah NH: **Pharmacovigilance using clinical notes**. *Clinical pharmacology and therapeutics* 2013, **93**(6):547-555.

28. Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH: **Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system**. *Clinical pharmacology and therapeutics* 2013, **93**(6):539-546.

29. Lee D, de Keizer N, Lau F, Cornet R: **Literature review of SNOMED CT use**. *Journal of the American Medical Informatics Association : JAMIA* 2014, **21**(e1):e11-19.

30. Nguyen A, Moore J, Zuccon G, Lawley M, Colquist S: **Classification of pathology reports for cancer registry notifications**. *Studies in health technology and informatics* 2012, **178**:150-156.

31. Villanueva-Rosales N, Dumontier M: **yOWL: an ontology-driven knowledge base for yeast biologists**. *J Biomed Inform* 2008, **41**(5):779-789.

32. Dumontier M, Villanueva-Rosales N: **Towards pharmacogenomics knowledge discovery with the semantic web**. *Briefings in Bioinformatics* 2009, **10**(2):153-163.

33. Cruz-Toledo J, Dumontier M, Parisien M, Major F: **RKB: a Semantic Web knowledge base for RNA**. *J Biomed Semantics* 2010, **1 Suppl 1**:S2.

34. Karp PD: **Artificial intelligence methods for theory representation and hypothesis formation**. *Comput Appl Biosci* 1991, **7**(3):301-308.

35. Karp PD: **Design methods for scientific hypothesis formation and their application to molecular biology**. *Machine Learning* 1993, **12**(1-3):89-116.

36. Karp PD: **HYPOTHESIS FORMATION AND QUALITATIVE REASONING IN MOLECULAR BIOLOGY**. Stanford University; 1989.

37. Karp PD, Ouzounis C, Paley S: **HinCyc: a knowledge base of the complete genome and metabolic pathways of H. influenzae**. *Proc Int Conf Intell Syst Mol Biol* 1996, **4**:116-124.

38. Karp PD, Riley M, Paley SM, Pelligrini-Toole A: **EcoCyc: an encyclopedia of Escherichia coli genes and metabolism**. *Nucleic Acids Research* 1996, **24**(1):32-39.

39. Zupan B, Bratko I, Demsar J, Juvan P, Curk T, Borstnik U, Beck JR, Halter J, Kuspa A, Shaulsky G: **GenePath: a system for inference of genetic networks and proposal of genetic experiments**. *Artificial intelligence in medicine* 2003, **29**(1-2):107-130.

40. Miller RA, Pople HE, Jr., Myers JD: **Internist-1, an experimental computer-based diagnostic consultant for general internal medicine**. *The New England journal of medicine* 1982, **307**(8):468-476.

41. Wolfram DA: **An appraisal of INTERNIST-I**. *Artificial intelligence in medicine* 1995, **7**(2):93-116.

42. Hayes-Roth F: **Rule-based systems**. *Communications of the ACM* 1985, **28**(9):921-932.

43. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A: **GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles**. *Bioinformatics* 2001, **17**(Suppl 1):S74-S82.

44. Danos V, Feret J, Fontana W, Harmer R, Krivine J: **Rule-Based Modelling of Cellular Signalling**. 2007, **4703**:17-41.

45. Hlavacek WS, Faeder JR, Blinov ML, Posner RG, Hucka M, Fontana W: **Rules for Modeling Signal-Transduction Systems**. *Science Signaling* 2006, **2006**(344):re6-re6.

46. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH: **Literature mining and database annotation of protein phosphorylation using a rule-based system**. *Bioinformatics* 2005, **21**(11):2759-2765.

47. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J: **ProMiner: rule-based protein and gene entity recognition**. *BMC Bioinformatics* 2005, **6 Suppl 1**:S14.

48. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova L *et al*: **The automation of science**. *Science* 2009, **324**(5923):85-89.

49. Soldatova L, King RD: **Representation of research hypotheses**. In: *Bio-Ontologies 2010: Semantic Applications in Life Sciences; Boston, MA*. 2010.

50. Tari L, Anwar S, Liang S, Cai J, Baral C: **Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism**. *Bioinformatics* 2010, **26**(18):i547-553.

51. Chen RO, Felciano R, Altman RB: **RIBOWEB: linking structural computations to a knowledge base of published experimental data**. *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:84-87.

52. Altman R, Bada M, Chai XJ, Whirl Carillo M, Chen RO, Abernethy N: **RiboWeb: An Ontology-Based System for Collaborative Molecular Biology**. *IEEE Intelligent Systems* 1999, **14**(5):68-76.

53. Racunas SA, Shah NH, Albert I, Fedoroff NV: **HyBrow: a prototype system for computer-aided hypothesis evaluation**. *Bioinformatics* 2004, **20**(suppl_1):i257-264.

54. Racunas SA, Shah NH, Fedoroff NV: **A case study in pathway knowledgebase verification**. *BMC Bioinformatics* 2006, **7**:196.

55. Langley P, Hunt G: **A Web-Based Environment for Explanatory Biological Modeling**. In: *AAAI Fall Symposium Series.* 2012.

56. Callahan A, Dumontier M, Shah NH: **HyQue: evaluating hypotheses using Semantic Web technologies**. *Journal of Biomedical Semantics* 2011, **2 Suppl 2**:S3.

57. Bhat PJ, Murthy TV: **Transcriptional control of the GAL/MEL regulon of yeast Saccharomyces cerevisiae: mechanism of galactose-mediated signal transduction**. *Mol Microbiol* 2001, **40**(5):1059-1066.

58. Kobayashi N, Ishii M, Takahashi S, Mochizuki Y, Matsushima A, Toyoda T: **Semantic-JSON: a lightweight web service interface for Semantic Web contents integrating multiple life science databases**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W533-540.

59. Callahan A, Dumontier M, Shah N: **HyQue: Evaluating hypotheses using Semantic Web technologies**. In: *Bio-Ontologies: Semantic applications in the life sciences: July 9-10 2010; Boston MA*. 2010.

60. Polikoff I: **Comparing SPIN with RIF** [http://topquadrantblog.blogspot.com/2011/06/comparing-spin-with-rif.html].

61. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**(5518):929-934.

62. Hoehndorf R, Dumontier M, Oellrich A, Rebholz-Schuhmann D, Schofield PN, Gkoutos GV: **Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning**. *PLoS One* 2011, **6**(7):e22006.

63. Berners-Lee T: **Linked Data** [http://www.w3.org/DesignIssues/LinkedData.html].

64. Nolin MA, Dumontier M, Belleau F, Corbeil J: **Building an HIV data mashup using Bio2RDF**. *Briefings in Bioinformatics* 2011.

65. Blonde W, Mironov V, Venkatesan A, Antezana E, De Baets B, Kuiper M: **Reasoning with bio-ontologies: using relational closure rules to enable practical querying**. *Bioinformatics* 2011, **27**(11):1562-1568.

66. Peng G, Hopper JE: **Evidence for Gal3p's cytoplasmic location and Gal80p's dual cytoplasmic-nuclear location implicates new mechanisms for controlling Gal4p activity in Saccharomyces cerevisiae**. *Molecular Cell Biology* 2000, **20**(14):5140-5148.

67. Goble C, Stevens R: **State of the nation in data integration for bioinformatics**. *J Biomed Inform* 2008, **41**(5):687-693.

68. Cerami EG, Bader GD, Gross BE, Sander C: **cPath: open source software for collecting, storing, and querying biological pathways**. *BMC Bioinformatics* 2006, **7**:497.

69. Chen H, Yu T, Chen JY: **Semantic Web meets Integrative Biology: a survey**. *Brief Bioinform* 2012.

70. Ruebenacker O, Moraru, II, Schaff JC, Blinov ML: **Integrating BioPAX pathway knowledge with SBML models**. *IET Syst Biol* 2009, **3**(5):317-328.

71. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L *et al*: **Toward interoperable bioscience data**. *Nat Genet* 2012, **44**(2):121-126.

72. Berlanga R, Jimenez-Ruiz E, Nebot V: **Exploring and linking biomedical resources through multidimensional semantic spaces**. *BMC bioinformatics* 2012, **13 Suppl 1**:S6.

73. Gennari JH, Neal ML, Galdzicki M, Cook DL: **Multiple ontologies in action: composite annotations for biosimulation models**. *Journal of Biomedical Informatics* 2011, **44**(1):146-154.

74. Hoehndorf R, Dumontier M, Gennari JH, Wimalaratne S, de Bono B, Cook DL, Gkoutos GV: **Integrating systems biology models and biomedical ontologies**. *BMC Syst Biol* 2011, **5**:124.

75. Jonquet C, Lependu P, Falconer S, Coulet A, Noy NF, Musen MA, Shah NH: **NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources**. *Web Semant* 2011, **9**(3):316-324.

76. Ruttenberg A, Rees JA, Samwald M, Marshall MS: **Life sciences on the Semantic Web: the Neurocommons and beyond**. *Brief Bioinform* 2009, **10**(2):193-204.

77. Momtchev V., Peychev D., Primov T., Georgiev G.: **Expanding the Pathway and Interaction Knowledge inLinked Life Data**. In: *Semantic Web Challenge: 2009; Amsterdam*.

78. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, Wild DJ: **Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data**. *BMC bioinformatics* 2010, **11**:255.

79. Campinas S, Perry TE, Ceccarelli D, Delbru R, Tummarello G: **Introducing RDF Graph Summary With Application to Assisted SPARQL Formulation**.

In: *23rd International Workshop on Database and Expert Systems Applications: September 3-7, 2012; Vienna, Austria*. 2012.

80. Ansell P: **Model and prototype for querying multiple linked scientific datasets**. *Future Generation Computer Systems* 2011, **27**(3):329-333.

81. Juty N, Le Novere N, Laibe C: **Identifiers.org and MIRIAM Registry: community resources to provide persistent identification**. *Nucleic Acids Res* 2012, **40**(Database issue):D580-586.

82. Mons B, van Haagen H, Chichester C, Hoen PB, den Dunnen JT, van Ommen G, van Mulligen E, Singh B, Hooft R, Roos M *et al*: **The value of data**. *Nat Genet* 2011, **43**(4):281-283.

83. Bizer C, Heath T, Berners-Lee T: **Linked Data - The Story So Far**. *International Journal on Semantic Web and Information Systems* 2009, **5**(3):1-22.

84. Zhao J, Sahoo SS, Missier P, Sheth A, Goble C: **Extending Semantic Provenance into the Web of Data**. *IEEE Internet Computing* 2011, **15**(1):40-48.

85. Moreau L, Clifford B, Freire J, Futrelle J, Gil Y, Groth P, Kwasnikowska N, Miles S, Missier P, Myers J *et al*: **The Open Provenance Model core specification (v1.1)**. *Future Generation Computer Systems* 2011, **27**(6):743-756.

86. Kuhn T, Barbano PE, Nagy ML, Krauthammer M: **Broadening the scope of nanopublications**. In: *Extended Semantic Web Conference: May, 2013; Montpellier, France*. 2013.

87. Groth P, Gibson A, Velterop J: **The anatomy of a nanopublication**. *Information Services & Use* 2010, **30**:51-56.

88. Patrinos GP, Cooper DN, van Mulligen E, Gkantouna V, Tzimas G, Tatum Z, Schultes E, Roos M, Mons B: **Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain**. *Hum Mutat* 2012, **33**(11):1503-1512.

89. Giardine B, Borg J, Higgs DR, Peterson KR, Philipsen S, Maglott D, Singleton BK, Anstee DJ, Basak AN, Clark B *et al*: **Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach**. *Nat Genet* 2011, **43**(4):295-301.

90. Schultes E, Chichester C, Burger K, Kotoulas S, Loizou A, Tkachenko V, Waagmeester A, Askjaer S, Pettifer S, Harland L *et al*: **The Open PHACTS Nanopublications Guidelines** [http://www.nanopub.org/guidelines/OpenPHACTS_Nanopublication_Guidlines_v1.8.1.pdf].

91. Clark T, Ciccarese PN, Goble C: **Micropublications: a Semantic Model for Claims, Evidence, Arguments and Annotations in Biomedical Communications** arXiv:1305.3506 [http://arxiv.org/abs/1305.3506].

92. Kuhn T, Dumontier M: **Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data**. In: *ESWC 2014: May 25-29th, 2014; Crete, Greece*. 2014.

93. Buneman P, Khanna S, Tan W-C: **Data Provenance: Some Basic Issues**. In: *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*. Edited by Kapoor S, Prasad S, vol. 1974: Springer Berlin Heidelberg; 2000: 87-93.

94. Bose R, Frew J: **Lineage retrieval for scientific data processing: a survey**. *ACM Computing Surveys* 2005, **37**(1):1-28.

95. Artz D, Gil Y: **A survey of trust in computer science and the Semantic Web**. *Web Semantics: Science, Services and Agents on the World Wide Web* 2007, **5**(2):58-71.

96. Grau BC, Horrocks I, Motik B, Parsia B, Patel-Schneider P, Sattler U: **OWL 2: The next step for OWL**. *Web Semantics: Science, Services and Agents on the World Wide Web* 2008, **6**(4):309-322.

97. Gibson A, van Dam J, Schultes E, Roos M, Mons B: **Towards computational evaluation of evidence for scientific assertions with nanopublications**. In: *Semantic Web Applications and Tools for Life Sciences 2012: November 28-30, 2012 2012; Paris, France*.

98. W3C: **RDF 1.1 Concepts and Abstract Syntax** [http://www.w3.org/TR/rdf11-concepts/].

99. Carroll JJ, Bizer C, Hayes P, Stickler P: **Named graphs, provenance and trust**. In: *Proceedings of the 14th international conference on World Wide Web: 2005*. ACM: 613-622.

100. W3C: **RDF Schema 1.1** [http://www.w3.org/TR/2014/REC-rdf-schema-20140225/].

101. W3C: **W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes** [http://www.w3.org/TR/2012/REC-xmlschema11-2-20120405/].

102. DCMI: **Dublin Core Metadata Element Set, Version 1.1** [http://dublincore.org/documents/2012/06/14/dces/].

103. Dumontier M, Baker CJO, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, Del Rio NR, Duck G, Furlong LI, Keath N *et al*: **The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery**. *Journal of Biomedical Semantics* 2014, **5**(1):14.

104. Razick S, Magklaras G, Donaldson IM: **iRefIndex: a consolidated protein interaction database with provenance**. *BMC bioinformatics* 2008, **9**:405.

105. Wittmeyer J, Joss L, Formosa T: **Spt16 and Pob3 of Saccharomyces cerevisiae form an essential, abundant heterodimer that is nuclear, chromatin-associated, and copurifies with DNA polymerase alpha**. *Biochemistry* 1999, **38**(28):8961-8971.

106. Orphanides G, LeRoy G, Chang CH, Luse DS, Reinberg D: **FACT, a factor that facilitates transcript elongation through nucleosomes**. *Cell* 1998, **92**(1):105-116.

107. Phan L, Zhang X, Asano K, Anderson J, Vornlocher HP, Greenberg JR, Qin J, Hinnebusch AG: **Identification of a translation initiation factor 3 (eIF3) core complex, conserved in yeast and mammals, that interacts with eIF5**. *Molecular and cellular biology* 1998, **18**(8):4935-4946.

108. Keener J, Dodd JA, Lalo D, Nomura M: **Histones H3 and H4 are components of upstream activation factor required for the high-level transcription of yeast rDNA by RNA polymerase I**. *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94**(25):13458-13462.

109. Kee Y, Lyon N, Huibregtse JM: **The Rsp5 ubiquitin ligase is coupled to and antagonized by the Ubp2 deubiquitinating enzyme**. *The EMBO journal* 2005, **24**(13):2414-2424.

110. McPhillips CC, Hyle JW, Reines D: **Detection of the mycophenolate-inhibited form of IMP dehydrogenase in vivo**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(33):12171-12176.

111. Hedstrom L: **IMP dehydrogenase: mechanism of action and inhibition**. *Current medicinal chemistry* 1999, **6**(7):545-560.

112. Fujitsu Laboratories Ltd.: **Fujitsu and DERI Revolutionize Access to Open Data by Jointly Developing Technology for Linked Open Data** [http://www.fujitsu.com/global/news/pr/archives/month/2013/20130403-02.html].

113. Urbani J, Kotoulas S, Maassen J, Van Harmelen F, Bal H: **WebPIE: A Web-scale Parallel Inference Engine using MapReduce**. *Web Semantics: Science, Services and Agents on the World Wide Web* 2012, **10**:59-75.

114. Kenyon CJ: **The genetics of ageing**. *Nature* 2010, **464**(7288):504-512.

115. Willcox BJ, Donlon TA, He Q, Chen R, Grove JS, Yano K, Masaki KH, Willcox DC, Rodriguez B, Curb JD: **FOXO3A genotype is strongly associated with human longevity**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(37):13987-13992.

116. Pawlikowska L, Hu D, Huntsman S, Sung A, Chu C, Chen J, Joyner AH, Schork NJ, Hsueh WC, Reiner AP *et al*: **Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity**. *Aging cell* 2009, **8**(4):460-472.

117. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R: **A C. elegans mutant that lives twice as long as wild type**. *Nature* 1993, **366**(6454):461-464.

118. Demetrius L: **Of mice and men. When it comes to studying ageing and the means to slow it down, mice are not just small humans**. *EMBO reports* 2005, **6 Spec No**:S39-44.

119. Rodriguez M, Snoek LB, De Bono M, Kammenga JE: **Worms under stress: C. elegans stress response and its relevance to complex human disease and aging**. *Trends in genetics : TIG* 2013, **29**(6):367-374.

120. Mair W, Dillin A: **Aging and survival: the genetics of life span extension by dietary restriction**. *Annual review of biochemistry* 2008, **77**:727-754.

121. Tatar M, Khazaeli AA, Curtsinger JW: **Chaperoning extended life**. *Nature* 1997, **390**(6655):30.

122. Hsu AL, Murphy CT, Kenyon C: **Regulation of aging and age-related disease by DAF-16 and heat-shock factor**. *Science* 2003, **300**(5622):1142-1145.

123. Ludewig AH, Izrayelit Y, Park D, Malik RU, Zimmermann A, Mahanti P, Fox BW, Bethke A, Doering F, Riddle DL *et al*: **Pheromone sensing regulates Caenorhabditis elegans lifespan and stress resistance via the deacetylase SIR-2.1**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(14):5522-5527.

124. Halaschek-Wiener J, Khattra JS, McKay S, Pouzyrev A, Stott JM, Yang GS, Holt RA, Jones SJ, Marra MA, Brooks-Wilson AR *et al*: **Analysis of long-lived C. elegans daf-2 mutants using serial analysis of gene expression**. *Genome research* 2005, **15**(5):603-615.

125. Plank M, Wuttke D, van Dam S, Clarke SA, de Magalhaes JP: **A meta-analysis of caloric restriction gene expression profiles to infer common signatures and regulatory mechanisms**. *Molecular bioSystems* 2012, **8**(4):1339-1349.

126. Wuttke D, Connor R, Vora C, Craig T, Li Y, Wood S, Vasieva O, Shmookler Reis R, Tang F, de Magalhaes JP: **Dissecting the gene network of dietary restriction to identify evolutionarily conserved pathways and new functional genes**. *PLoS genetics* 2012, **8**(8):e1002834.

127. Ludewig AH, Klapper M, Doring F: **Identifying evolutionarily conserved genes in the dietary restriction response using bioinformatics and subsequent testing in Caenorhabditis elegans**. *Genes & nutrition* 2014, **9**(1):363.

128. Wang Z, Sagotsky J, Taylor T, Shironoshita P, Deisboeck TS: **Accelerating cancer systems biology research through Semantic Web technology**. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2013, **5**(2):135-151.

129. Harrow I, Filsell W, Woollard P, Dix I, Braxenthaler M, Gedye R, Hoole D, Kidd R, Wilson J, Rebholz-Schuhmann D: **Towards virtual knowledge broker services for semantic integration of life science literature and data sources**. *Drug discovery today* 2013, **18**(9-10):428-434.

130. Hancock JM: **Editorial: biological ontologies and semantic biology**. *Frontiers in genetics* 2014, **5**:18.

131. Croset S, Overington JP, Rebholz-Schuhmann D: **The functional therapeutic chemical classification system**. *Bioinformatics* 2013.

132. Callahan A, Dumontier M, Shah NH: **HyQue: evaluating hypotheses using Semantic Web technologies**. *J Biomed Semantics* 2011, **2 Suppl 2**:S3.

133. Callahan A, Dumontier M: **Evaluating Scientific Hypotheses Using the SPARQL Inferencing Notation**. In: *The Semantic Web: Research and Applications.* Edited by Simperl E, Cimiano P, Polleres A, Corcho O, Presutti V, vol. 7295: Springer; 2012: 647-658.

134. Knublauch H, Hendler JA, Idehen K: **SPIN - Overview and Motivation** [http://www.w3.org/Submission/spin-overview/].

135. Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhaes JP: **Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing**. *Nucleic acids research* 2013, **41**(Database issue):D1027-1033.

136. Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Done J, Grove C, Howe K *et al*: **WormBase 2014: new views of curated biology**. *Nucleic acids research* 2014, **42**(1):D789-793.

137. W3C: **RDF Primer - W3C Recommendation 10 February 2004** [http://www.w3.org/TR/2004/REC-rdf-primer-20040210/].

138. Callahan A, Cruz-Toledo J, Ansell P, Dumontier M: **Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data**. In: *The Semantic Web: Semantics and Big Data.* Edited by Cimiano P, Corcho O, Presutti V, Hollink L, Rudolph S, vol. 7882: Springer Berlin Heidelberg; 2013: 200-212.

139. Van Nostrand EL, Sanchez-Blanco A, Wu B, Nguyen A, Kim SK: **Roles of the developmental regulator unc-62/Homothorax in limiting longevity in Caenorhabditis elegans**. *PLoS genetics* 2013, **9**(2):e1003325.

140. Zarse K, Schmeisser S, Groth M, Priebe S, Beuster G, Kuhlow D, Guthke R, Platzer M, Kahn CR, Ristow M: **Impaired insulin/IGF1 signaling extends life span by promoting mitochondrial L-proline catabolism to induce a transient ROS signal**. *Cell metabolism* 2012, **15**(4):451-465.

141. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nature protocols* 2012, **7**(3):562-578.

142. Bodenreider O, Aubry M, Burgun A: **Non-lexical approaches to identifying associative relations in the gene ontology**. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2005:91-102.

143. Faria D, Schlicker A, Pesquita C, Bastos H, Ferreira AE, Albrecht M, Falcao AO: **Mining GO annotations for improving annotation consistency**. *PloS one* 2012, **7**(7):e40519.

144. Birnby DA, Link EM, Vowels JJ, Tian H, Colacurcio PL, Thomas JH: **A transmembrane guanylyl cyclase (DAF-11) and Hsp90 (DAF-21) regulate a common set of chemosensory behaviors in caenorhabditis elegans**. *Genetics* 2000, **155**(1):85-104.

145. Morley JF, Morimoto RI: **Regulation of longevity in Caenorhabditis elegans by heat shock factor and molecular chaperones**. *Molecular biology of the cell* 2004, **15**(2):657-664.

146. Yu H, Larsen PL: **DAF-16-dependent and independent expression targets of DAF-2 insulin receptor-like pathway in Caenorhabditis elegans include FKBPs**. *Journal of molecular biology* 2001, **314**(5):1017-1028.

147. Inada H, Ito H, Satterlee J, Sengupta P, Matsumoto K, Mori I: **Identification of guanylyl cyclases that function in thermosensory neurons of Caenorhabditis elegans**. *Genetics* 2006, **172**(4):2239-2252.

148. Murphy CT, McCarroll SA, Bargmann CI, Fraser A, Kamath RS, Ahringer J, Li H, Kenyon C: **Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans**. *Nature* 2003, **424**(6946):277-283.

149. Pujol N, Link EM, Liu LX, Kurz CL, Alloing G, Tan MW, Ray KP, Solari R, Johnson CD, Ewbank JJ: **A reverse genetic analysis of components of the Toll signaling pathway in Caenorhabditis elegans**. *Current biology : CB* 2001, **11**(11):809-821.

150. Boulton SJ, Gartner A, Reboul J, Vaglio P, Dyson N, Hill DE, Vidal M: **Combined functional genomic maps of the C. elegans DNA damage response**. *Science* 2002, **295**(5552):127-131.

151. Prufer K, Muetzel B, Do HH, Weiss G, Khaitovich P, Rahm E, Paabo S, Lachmann M, Enard W: **FUNC: a package for detecting significant associations between gene sets and ontological annotations**. *BMC bioinformatics* 2007, **8**:41.

152. Shaye DD, Greenwald I: **OrthoList: a compendium of C. elegans genes with human orthologs**. *PLoS One* 2011, **6**(5):e20085.

153. van der Walt JM, Noureddine MA, Kittappa R, Hauser MA, Scott WK, McKay R, Zhang F, Stajich JM, Fujiwara K, Scott BL *et al*: **Fibroblast growth factor 20 polymorphisms and haplotypes strongly influence risk of Parkinson disease**. *American journal of human genetics* 2004, **74**(6):1121-1127.

154. Satake W, Mizuta I, Suzuki S, Nakabayashi Y, Ito C, Watanabe M, Takeda A, Hasegawa K, Sakoda S, Yamamoto M *et al*: **Fibroblast growth factor 20 gene and Parkinson's disease in the Japanese population**. *Neuroreport* 2007, **18**(9):937-940.

155. Yuan Y, Tong Q, Zhou X, Zhang R, Qi Z, Zhang K: **The association between glycogen synthase kinase 3 beta polymorphisms and Parkinson's disease susceptibility: a meta-analysis**. *Gene* 2013, **524**(2):133-138.

156. Zhang N, Yu JT, Yang Y, Yang J, Zhang W, Tan L: **Association analysis of GSK3B and MAPT polymorphisms with Alzheimer's disease in Han Chinese**. *Brain research* 2011, **1391**:147-153.

157. Mondragon-Rodriguez S, Perry G, Zhu X, Moreira PI, Williams S: **Glycogen synthase kinase 3: a point of integration in Alzheimer's disease and a therapeutic target?** *International journal of Alzheimer's disease* 2012, **2012**:276803.

158. Morgan JC, Currie LJ, Harrison MB, Bennett JP, Jr., Trugman JM, Wooten GF: **Mortality in levodopa-treated Parkinson's disease**. *Parkinson's disease* 2014, **2014**:426976.

159. Rait G, Walters K, Bottomley C, Petersen I, Iliffe S, Nazareth I: **Survival of people with clinical diagnosis of dementia in primary care: cohort study**. *BMJ* 2010, **341**:c3584.

160. Rowland LP, Shneider NA: **Amyotrophic lateral sclerosis**. *The New England journal of medicine* 2001, **344**(22):1688-1700.

161. Judson PN, Stalford SA, Vessey J: **Assessing confidence in predictions made by knowledge-based systems**. *Toxicology Research* 2013, **2**(1):70.

162. Hendler J, Berners-Lee T: **From the Semantic Web to social machines: A research challenge for AI on the World Wide Web**. *Artificial Intelligence* 2010, **174**(2):156-161.

163. Blake JA, Bult CJ: **Beyond the data deluge: data integration and bio-ontologies**. *J Biomed Inform* 2006, **39**(3):314-320.

164. Gomez-Perez A, Martinez-Romero M, Rodriguez-Gonzalez A, Vazquez G, Vazquez-Naya JM: **Ontologies in medicinal chemistry: current status and future challenges**. *Current topics in medicinal chemistry* 2013, **13**(5):576-590.

165. Bird CL, Frey JG: **Chemical information matters: an e-Research perspective on information and data sharing in the chemical sciences**. *Chemical Society reviews* 2013, **42**(16):6754-6776.

166. Duerr JS: **Immunohistochemistry**. In: *WormBook.* Edited by The C. elegans Research Community; 2006.

167. Lee M-H, Schedl T: **RNA in situ hybridization of dissected gonads**. In: *WormBook.* Edited by Community TCeR; 2006.

168. Motohashi T, Tabara H, Kohara Y: **Protocols for large scale in situ hybridization on C. elegans larvae**. In: *WormBook*. Edited by Community TCeR; 2006.

169. Sutphin GL, Kaeberlein M: **Measuring Caenorhabditis elegans life span on solid media**. *Journal of visualized experiments : JoVE* 2009(27).

170. McNaughton R, Huet G, Shakir S: **An investigation into drug products withdrawn from the EU market between 2002 and 2011 for safety reasons and the evidence used to support the decision-making**. *BMJ open* 2014, **4**(1):e004221.

171. Force T, Kolaja KL: **Cardiotoxicity of kinase inhibitors: the prediction and translation of preclinical models to clinical outcomes**. *Nature reviews Drug discovery* 2011, **10**(2):111-126.

172. Wilkinson MD, McCarthy L, Vandervalk B, Withers D, Kawas E, Samadian S: **SADI, SHARE, and the in silico scientific method**. *BMC bioinformatics* 2010, **11 Suppl 12**:S7.

173. Chepelev LL, Dumontier M: **Semantic Web integration of Cheminformatics resources with the SADI framework**. *J Cheminform* 2011, **3**:16.

174. Chepelev LL, Riazanov A, Kouznetsov A, Low HS, Dumontier M, Baker CJ: **Prototype semantic infrastructure for automated small molecule classification and annotation in lipidomics**. *BMC bioinformatics* 2011, **12**:303.

175. Riazanov A, Laurila JB, Baker CJ: **Deploying mutation impact text-mining software with the SADI Semantic Web Services framework**. *BMC bioinformatics* 2011, **12 Suppl 4**:S6.

176. Domingos P: **A few useful things to know about machine learning**. *Communications of the ACM* 2012, **55**(10):78.

177. Halevy A, Norvig P, Pereira F: **The Unreasonable Effectiveness of Data**. *IEEE Intelligent Systems* 2009, **24**(2):8-12.