Structuring the life science resourceome for Semantic Systems Biology: lessons from the BioGateway project

Conference Paper · November 2008

CITATIONS
8

READS
281

8 authors, including:

Erick Antezana
Norwegian University of Science and Technology
40 PUBLICATIONS **863** CITATIONS
SEE PROFILE

Ward Blondé
Ghent University
39 PUBLICATIONS **254** CITATIONS
SEE PROFILE

Mikel Egaña
University of the Basque Country
39 PUBLICATIONS **899** CITATIONS
SEE PROFILE

Alistair Rutherford
Mott MacDonald Group
2 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Structuring the life sciences resourceome for Semantic Systems Biology: lessons from the BioGateway project

Erick Antezana[1,2], Ward Blondé[1,2], Mikel Egaña[3], Alistair Rutherford[4], Robert Stevens[3], Bernard De Baets[5], Vladimir Mironov[6], and Martin Kuiper[6]

[1] Dept. of Plant Systems Biology, VIB, Gent, Belgium
[2] Dept. of Molecular Genetics, Ghent University, Belgium
[3] School of Computer Science, The University of Manchester, UK
[4] http://www.netthreads.co.uk, Glasgow, Scotland, UK
[5] Dept. of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium
[6] Dept. of Biology, Norwegian University of Science and Technology, Norway
{erant|wablo}@psb.ugent.be
{eganaarm|stevensr}@cs.man.ac.uk
alistair.rutherford@gmail.com
bdebaets@ugent.be
{mironov|kuiper}@bio.ntnu.no

Abstract. The application of Semantic Web technologies in the life sciences for data integration is still nascent. We have recently built BioGateway, an RDF store that integrates all the candidate OBO Foundry ontologies with other resources such as SWISS-PROT. In the course of developing BioGateway, we faced challenges that are common to other projects that involve large datasets in diverse formats. We present a detailed analysis of the obstacles that had to be solved in creating BioGateway. In doing so, we demonstrate the potential of a comprehensive application of Semantic Web technologies to global biomedical data. The time is ripe for launching a community effort aiming at a wider acceptance and application of Semantic Web technologies in the life sciences domain. We make a public call for the creation of a forum that strives to implement a truly semantic life science foundation of a type of Systems Biology that we named Semantic Systems Biology.

1 Introduction

We witness a growing acceptance of Semantic Web technologies by the life science community for the purpose of knowledge management. This is illustrated by the existence of a W3C special interest group[7] (HCLS IG) and many other projects that exploit semantic technologies, such as the Resource Description Framework

[7] http://www.w3.org/2001/sw/hcls/

(RDF)[8] and the Web Ontology Language (OWL)[9], to represent biological information [1–7]. We are, however, just at the beginning of this process [8], and there are still many issues to be solved in order to build a semantic infrastructure that is adequate for biological knowledge management. Such an infrastructure will not only allow more efficient knowledge management; it will make possible a much more integrated and contextualised approach towards biomedical research. Semantic Web technologies have the potential to add a new dimension of knowledge integration to Systems Biology (SB), which is expected to be among the early adopters of these technologies [1]. We call this combination *Semantic Systems Biology* (SSB), a form of systems biology where new hypotheses about a biological system are not generated through a mathematical model but through global queries and reasoning on integrated data.

As part of our work towards SSB, we constructed BioGateway[10], a system built upon an RDF store that aggregates bio-ontologies and other bioinformatics resources. It provides protein information for all the species with annotated genomes. Data integration is an important component in an SB approach, and with BioGateway we add a semantic foundation. But more important for SB is mathematical modelling. By integrating a systems network with a mathematical model, one can simulate the behaviour of the network, and predict the outcome of new experiments. With semantic knowledge bases, a querying and reasoning component could be added to this, where a mathematical model is not exploited, but new hypotheses about the system and its components are obtained. In short, the paradigm of Semantic Systems Biology acts as a complement to "standard" Systems Biology.

BioGateway allows a bioinformatician or biologist to query across a semantically integrated collection of resources at a systems level. BioGateway illustrates both the challenges and the benefits that the Semantic Web brings to the life sciences, and we therefore elaborate in this paper on its technical properties and demonstrate its utility. Some of the problems we faced while building BioGateway lead us to conclude that there is a need for a wider Semantic Systems Biology forum to promote standards.

2 BioGateway data model

2.1 BioGateway graphs

BioGateway is a system holding an RDF store that combines information from different resources[11]: the entire set of candidate Open Biomedical Ontologies (OBO) Foundry ontologies [9], the complete collection of annotations provided by the Gene Ontology Annotation (GOA) files [10], a simplified version of the NCBI taxonomy [11] (including the names, ranks, and taxonomical hierarchy), a

[8] http://www.w3.org/RDF/
[9] http://www.w3.org/2004/OWL/
[10] http://www.semantic-systems-biology.org/biogateway
[11] http://www.semantic-systems-biology.org/biogateway/resources

subset of SWISS-PROT [12] (excluding the sequences themselves, for instance), and the Cell Cycle Ontology (CCO)[12].

All the imported data sources, when converted to RDF graphs, share a basic URI:

http://www.semantic-systems-biology.org

This means that each resource (*e.g.* each protein from SWISS-PROT, each taxon from the NCBI taxonomy, each OBO term) has a URI of the form:

http://www.semantic-systems-biology.org/SSB#resource

Each of the imported data sources is represented as an individual graph with a specific URI, of the following form:

http://www.semantic-systems-biology.org/graph_name

Additionally, the SSB graph combines all the constituent graphs of BioGateway, containing about 175 million triples. Intermediate graphs for the GOA files and the OBO Foundry candidate ontologies contain about 160 million triples and 8 million triples respectively.

Many of the RDF graphs in BioGateway contain orthogonal resources not connected to each other, like SWISS-PROT and the OBO Foundry ontologies. SWISS-PROT resources are, however, linked to GO resources via GOA resources. This also interlinks the three sub-ontologies of GO. To accommodate evidence codes from GOA, a reified or n-ary node is created. For example, the following excerpt from a GOA file[13] would be converted into the RDF structure shown in Figure 1:

```
UniProtKB O03042 O03042  GO:0000287 GOA:spkw|GO_REF:0000004 IEA
```

2.2 BioGateway scaffold: BioMetarel and MetaOnto

Two ontologies were created in order to provide a scaffold to integrate all the graphs: Metaonto and BioMetarel.

BioMetarel[14] holds the predicate types or relation types used to link subjects to objects. It also links the unique id's of the relation types with their user-friendly names. BioMetarel also contains all the meta-information, like transitivity and reflexivity, about the biomedical relation types that are used. This relation ontology consists of a generic scaffold, the Metarel ontology[15] , to which all the relation types of RO [13], and all the relation types that are used in the OBO Foundry ontologies, are added. Unfortunately, these relation types were

[12] http://www.cellcycleontology.org/
[13] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/3.A_thaliana.goa
[14] http://www.bioontology.org/files/38667/biometarel.obo
[15] http://www.semantic-systems-biology.org/metarel

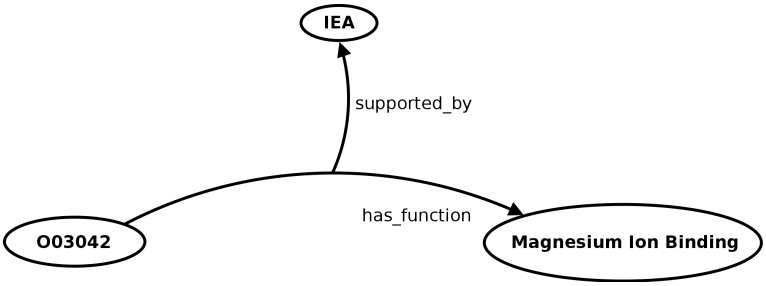

Fig. 1. RDF model of a GOA entry. The protein O03042 (Ribulose bisphosphate carboxylase large chain) is annotated with the GO term GO:0000287 (Magnesium Ion Binding), a term in the Molecular Function subtree from GO. Therefore, O03042 has the molecular function of binding magnesium ion. This fact is supported by IEA, that is, Inferred from Electronic Annotation.

not named consistently throughout the candidate ontologies (*e.g.* the subsumption relation was called both *is a* and *Is A*, and the partonomic relation both *part of* and *is part of*). A consistent list of relation types was manually created for BioMetarel. We chose as a rule to include a verb in every relation type name, conjugated as the third person singular in the present tense. The application of this rule predominantly involved the addition of the verb *is*. As a consequence, we can return triples in the form of a *pseudo*-grammatical sentence, like *blood is located in vein*. This rule also prompted us to transform names like *anatomical_relation* to *is anatomically related to* and *surrounding* to *surrounds*. The meaning of several poorly named relation types in fact became clearer by adhering to this format. The RDF file of BioMetarel is uploaded as a separate graph in BioGateway.

The most straight-forward use of BioMetarel is to connect the unique id's of the relation types with their user-friendly names. We observed, however, that the inclusion of the full BioMetarel interfered with some specific queries, like the listing of all the resources of a graph. Therefore, we created a lightweight subontology of BioMetarel, called Biorel. This subontology contains only relation types without the metaclasses and metarelations between relation types. This made Biorel more suited to be included in every single RDF graph in BioGateway.

Having such relations infrastructure implemented in BioGateway enabled us to build a consistent RDF scaffold for other resources such as evidence codes and GOA associations. All we needed to create the integrated graph was to consistently use appropriate identifiers for the predicates in the RDF triples. The integration of OBO Foundry ontologies with respect to the classes did not pose problems, because these get different identifiers in different ontologies, and they should be orthogonal as a design principle.

A small ontology, Metaonto, was created in the OBO format for the mapping between the names of the OBO ontologies and the prefixes they use in their unique id's. The mapping is very useful for users who want to explore the OBO

Foundry with queries in BioGateway. Meta-information like the names of the RDF graphs, the names of the OBO ontologies and characteristics of the relation types are accessible as results of the so called "ontological queries" (in opposition to "biological queries", see Section 5).

In summary, the integration of data in BioGateway has been achieved on the basis of the use of BioMetarel, the use of the same URIs for equivalent resources in the data sources (SwissProt, GOA, NCBI taxonomy) and the orthogonality of OBO ontologies with respect to the classes.

3 Design of BioGateway

While defining the specification of the RDF-translations for each of the integrated resources, we also developed a library of queries (see Section 5). This resulted in an RDF model that is adequately suited for querying, in particular in terms of performance. During this process we have paid attention to several quality constraints:

1. *Quick results:* A relatively quick query answer is always a desirable feature for any system. Therefore, we have systematically tested the response time with a suite of queries. This quality constraint turned out to be the biggest challenge during the development of the system. One extra line (triple) in a SPARQL[16] query could mean a huge difference for the computational performance. This is one of the reasons why we did not pick existing ontologies or Systems Biology resources represented in OWL's (RDF/XML syntax), such as BioPAX [6] or SBML[17]. The verbosity of OWL's RDF/XML might work satisfactorily in other query systems having small OWL models, but it is a heavy burden for efficient SPARQL querying using current solutions. We could, however, substantially reduce the length of queries by RDF optimisation. As BioGateway was increasing in size during its development, the computational performance was decreasing dramatically when new resources were integrated. Therefore, next to the single graph integrating all the resources (SSB graph), we created RDF graphs corresponding to each of the constituent resources of BioGateway, which can still be combined in queries. By these optimisations, many queries answer within a second, while others can require about 10 seconds.
2. *Human readable output:* As RDF works with URIs, many outputs from SPARQL queries might be hard to comprehend. We tried to avoid such outputs as much as possible by creating labels for all the terms and all the relation types. These can be used to present the results to the user.
3. *Good practice:* RDF is a Semantic Web standard that implies good design practices[18] when it comes to integration with other efforts within the framework of the Semantic Web. Orthogonality was achieved for all the terms,

[16] http://www.w3.org/TR/rdf-sparql-query/
[17] http://sbml.org/Documents/Specifications
[18] http://www.w3.org/TR/2008/WD-swbp-vocab-pub-20080123/

meaning that the proteins in SWISS-PROT received the same unique id's as the proteins in GOA. Combining these graphs in a single query would not otherwise be possible.

3.1 Simulating transitive closure

Transitive closure is an important feature in biomedical knowledge representation, especially where it concerns partonomy [14]. In addition, transitive closure along the *is_a* relation type is also desirable. Transitivity, however, cannot be expressed in RDF, and therefore it had to be created explicitly by adding all the necessary triples programmatically when loading the resources into the RDF triple store dedicated to BioGateway (see Section 3.2). That is, if resources A, B and C are related via *part_of* (A *part_of* B *part_of* C), a third triple A *part_of* C is created. This operation was done for the candidate OBO ontologies, the Cell Cycle Ontology and BioMetarel allowing transitivity in queries to be exploited with little impact on the performance of BioGateway. The ONTO-PERL [15] utility, used for adding the transitive closure over is_a and part_of, can be customized to consider other types of relations (e.g. located_in).

3.2 BioGateway architecture

BioGateway serves as a gateway to distributed resources on the Web. An automated pipeline downloads the latest released resources on a local server every two months. The majority of the downloaded resources are converted to RDF using the ONTO-PERL suite [15], which contains RDF converters for the following formats: the OBO files (OBO format[19]), the tab delimited GOA files[20], the NCBI taxonomy dump files[21] and the SWISS-PROT entry files [12]. In addition, ONTO-PERL generates the necessary transitive closure graphs (see Section 3.1).

After that, the RDF files are uploaded into RDF graphs in Open Virtuoso[22], which contains an endpoint (http://crunch.fvms.ugent.be:8891/sparql), where SPARQL queries can be submitted. A user interface (http://www.semantic-systems-biology.org/biogateway/querying) with a library of queries and an edit-box points to the SPARQL endpoint, through simple HTML-technology (Figure 2).

4 Visualisation of query results

The visualisation of triple-based resources poses a special challenge. It is necessary to develop and deploy new interfaces to manipulate, query and visualize this knowledge in an intuitive way. A SPARQL browser (still under development)

[19] http://www.geneontology.org/GO.format.obo-1_2.shtml
[20] http://www.ebi.ac.uk/GOA/goaHelp.html#4
[21] ftp://ftp.ncbi.nih.gov/pub/taxonomy/
[22] http://virtuoso.openlinksw.com/

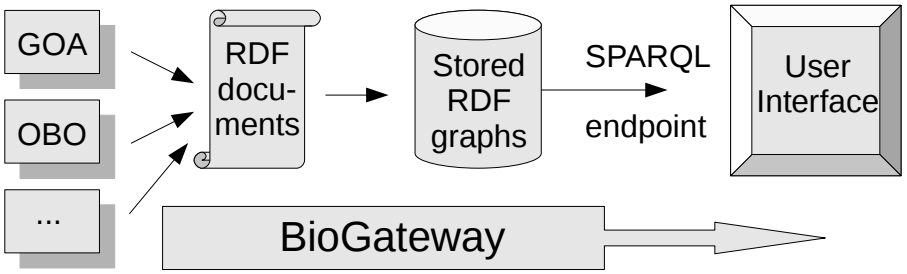

Fig. 2. Information resources are converted to RDF documents that are uploaded to a triple store (OpenVirtuoso), where they can be queried using SPARQL.

enables querying and visual exploration of the results obtained using the BioGateway. It can be accessed from the SSB website[23]. With this interface, users can define a SPARQL query over BioGateway resources, the SPARQL endpoint could also be customised (by default it points to the SSB endpoint[24]) (Figure 3). After executing a query, a network of results is displayed (Figure 4). A tabular representation of the result is also available. The SPARQL browser has been developed using Flex technologies[25], which provide powerful ways of creating interfaces with dynamic features. The entire source code is freely available[26].

5 Queries

SPARQL queries can be executed against the BioGateway triple store[27]. Many sample queries are available at the web site, for example, the query in Figure 5 returns all the human proteins that are located in the nucleus (note the use of transitivity).

5.1 One-click query access

BioGateway provides a library of optimized, easily customisable SPARQL queries that make the resources easily accessible to both layman users and experts. Even SPARQL experts will not easily find their way through RDF resources with which they are not acquainted. Therefore, we tried to reflect the basic query requirements in the library. It makes BioGateway accessible with a single click and it is a building block for future applications.

The library was split into a section with *biological queries* and a section with *ontological queries*. The *biological queries* are designed for usage by biomedical

[23] http://www.semantic-systems-biology.org/sparql-viewer
[24] http://crunch.fvms.ugent.be:8891/sparql
[25] http://www.adobe.com/products/flex/
[26] http://www.netthreads.co.uk
[27] http://www.semantic-systems-biology.org/biogateway/querying

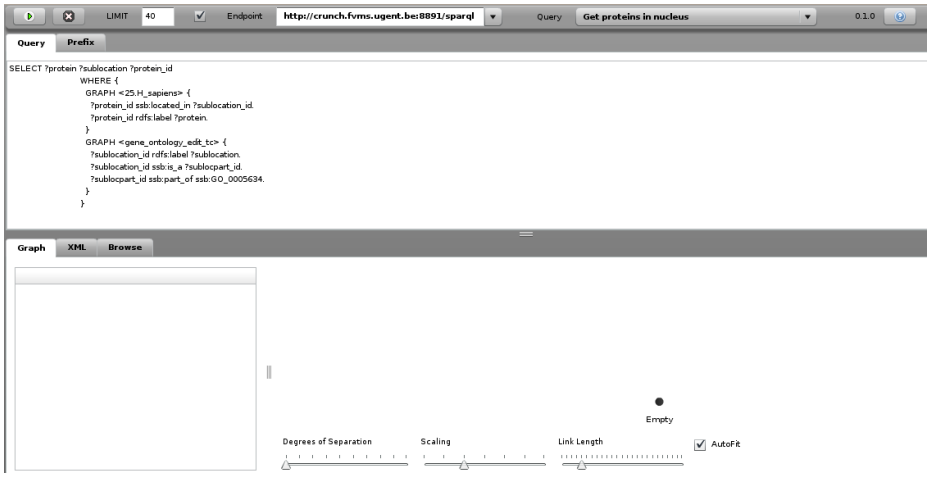

Fig. 3. SPARQL viewer. The queries are selected from the drop-down menu on the top right: in this case, the query "Get proteins in the nucleus" is selected. Queries can be customised, for example, by changing the parameters.

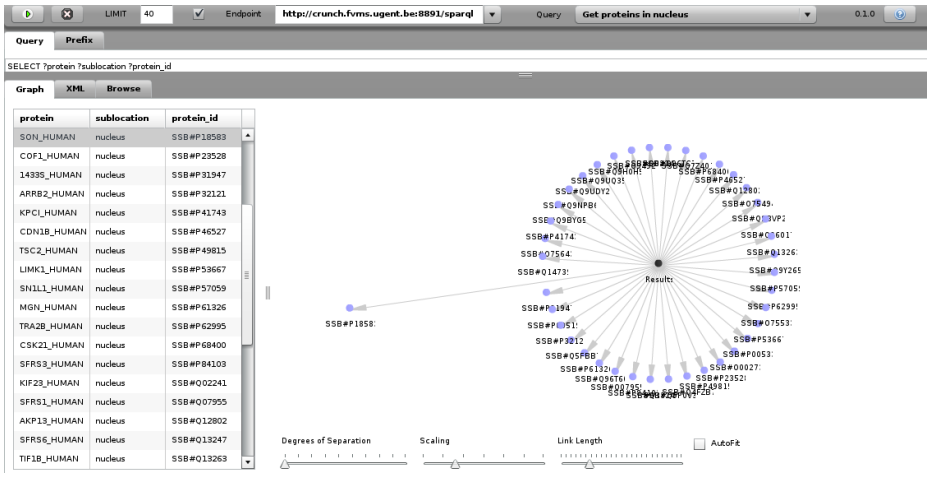

Fig. 4. SPARQL viewer. The query from Figure 3 has been executed, and the results are displayed. The appearance of the network can be configured.

```
# NAME     : get_proteins_in_nucleus
# PARAMETER: GO_0005634: the nucleus
# PARAMETER: 25.H_sapiens: the GOA graph for human
# FUNCTION : returns all the human proteins that have the
#            nucleus as annotated location

BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
SELECT ?protein ?sublocation ?protein_id
WHERE {
  GRAPH <25.H_sapiens> {
    ?protein_id ssb:located_in ?sublocation_id.
    ?protein_id rdfs:label ?protein.
  }
  GRAPH <gene_ontology_edit_tc> {
    ?sublocation_id rdfs:label ?sublocation.
    ?sublocation_id ssb:is_a ?sublocpart_id.
    ?sublocpart_id ssb:part_of ssb:GO_0005634.
  }
}
```

Fig. 5. SPARQL query sample to retrieve human proteins that are located in the nucleus. The metadata about the query are presented in the first 5 lines on the top: name, parameters that can be changed and function. The rest constitutes the query itself.

scientists, and they draw on the most relevant part of the knowledge base. Some examples of biological queries read as follows:

1. Get the proteins with a specific function/location/process for any of the annotated organisms. For example in Figure 5 a query that returns all the human proteins that are located in the nucleus can be seen.
2. Get the information on the function, location, process and associated disease for a given protein.
3. Get the proteins that are involved in the "psoriasis" disease.

On the other hand, the set of *ontological queries* shows how SPARQL can be used to explore BioGateway and specifically the OBO ontologies. This set of queries is intended for users interested in ontology engineering. Any future applications that build on the results of SPARQL queries will certainly benefit from the availability of basic navigation-type queries like *get neighborhood*, *get the root of an ontology*, *get the hierarchy to the root*, *get graphs*, *etc.* These queries explore the typical network structure of RDF models. On the other hand, the ontological queries show the RDF semantics that are available in BioGateway, like subsumption, transitivity and composition of relations. Some examples of ontological queries read as follows:

1. Query the OBO Foundry: search on names and get their unique id's.
2. Get all the neighbor terms of a given term.
3. Get all the properties, like definition, synonyms, *etc.*, of a given OBO term.

Both sections of the library make BioGateway a workbench for creating SPARQL queries. Often, the results of a query can be used to copy-paste as a parameter in other queries. We elaborate this idea further in Section 5.2.

All the queries in the library were provided with a name, their function and a list of parameters that can be customised in a query. By using prefixes properly, a SPARQL query can be written in such a way that a parameter needs replacement only in one fixed place. All the queries in the library were written in that way.

5.2 Combining regular RDF graphs with transitive closure graphs.

One of the ontological queries in the library is designed to find the closest common ancestor in the hierarchy of an ontology for two given terms (Figure 6).

For this query we need both the regular RDF ontology and its transitive closure (SSB_tc, which is generated by the pipeline, see Section 3.2) . In fact, the query might be reduced to: *find all the ancestors of both terms that do not have any descendants that are ancestral to both terms.* To find all the terms that are ancestors of both terms, we need the transitive closure graph, as in that form all the ancestors are directly linked to their descendants. Two triples in the query are enough to retrieve their id:

```
GRAPH <SSB_tc> {
term1_id: ssb:is_a ?common_ancestor_id.
term2_id: ssb:is_a ?common_ancestor_id.
}
```

```
# NAME     : get_common_ancestor
# PARAMETER: GO_0002617: the first query-term
# PARAMETER: GO_0034125: the second query-term
# FUNCTION : returns the closest common ancestor-term in the
#            hierarchy for two given terms
BASE   <http://www.semantic-systems-biology.org/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#>
PREFIX term1_id: <SSB#GO_0002617>
PREFIX term2_id: <SSB#GO_0034125>
SELECT distinct ?common_ancestor ?common_ancestor_id
WHERE {
  GRAPH <SSB_tc> {
    term1_id: ssb:is_a ?common_ancestor_id.
    term2_id: ssb:is_a ?common_ancestor_id.
    OPTIONAL {
      term1_id: ssb:is_a ?direct_child.
      term2_id: ssb:is_a ?direct_child.
      GRAPH <SSB> {
        ?direct_child ssb:is_a ?common_ancestor_id.
      }
    }
    ?common_ancestor_id rdfs:label ?common_ancestor.
  }
  FILTER(!bound(?direct_child))
}
```

Fig. 6. Getting the closest common ancestor of 2 terms.

We get all common ancestors with this query, while we only want the closest ones. Therefore, we check for the children of this set of ancestors. This can be best accomplished in the ontology without transitive closure:

```
GRAPH <SSB> {
?direct_child ssb:is_a ?common_ancestor_id.
}
```

Additionally, we check whether these children belong to the same set of common ancestors as defined before:

```
term1_id: ssb:is_a ?direct_child.
term2_id: ssb:is_a ?direct_child.
```

The last two checks go in an optional clause, because we only want the common ancestors for which these checks fail. In this way, we can filter the common ancestors for which this kind of *?direct_child* does not exist:

```
FILTER(!bound(?direct_child))
```

6 Discussion

Currently, the life sciences community is becoming aware of the need for standards and standardised ways to archive data and metadata [16, 17, 9]. The W3C provides formal standards to represent knowledge (*e.g.* RDF, OWL). Although there are still limitations with respect to the representation of some type of information (e.g. spatio-temporal information) or it may prove difficult to model complex scenarios (such as expression data from microarray experiments), these standards have been shown to accommodate information that can be queried to gain further biological insights [18–20].

Here we have presented an RDF triple store, named BioGateway, that integrates different life science knowledge resources. Other projects [21–24, 4] have attempted similar integration. They, however, either used smaller sets of data or offered limited query possibilities due to performance issues.

In the future, BioGateway will include more resources, for example OBO cross products[28]. Other BioGateway extensions will include an improved SPARQL interface and an enhanced integration of graphs.

In an ideal world where the Semantic Web is completely operational, BioGateway would probably be obsolete. That ideal world is, however, still far away, and many issues will have to be solved if the Semantic Web is ever to become fully operational.

[28] http://obofoundry.org/index.cgi?show=mappings

Biological identifiers A universal resolvable mechanism for identifying biological entities is vital for a life sciences Semantic Web [8]. There have been different attempts in that direction. For example, the OBO Foundry insists that the ontologies have unique identifiers that are orthogonal to identifiers in other OBO Foundry ontologies. Such identifiers, however, are not resolvable and therefore not scalable [8]. Currently, there are mechanisms proposed for resolvable identifiers, such as URIs[29], LSIDs[30], OKKAM IDs[31] and MIRIAM URIs [25]. URIs are used for identifiers in OWL or RDF ontologies, and therefore they offer an efficient foundation. In the case of BioGateway, *ad hoc* URL-based identifiers were created. They are expected to be standardised once the SSB forum is established.

Lack of semantic content Most biological information is either not adequately semantically codified or it has been codified with a poor axiomatisation [26]. This information should be richly codified using Semantic Web languages like RDF or OWL, which is not a trivial task given the disparity of the assumptions behind languages like OBOF or OWL [27, 28].

Most biologists are still unaware of the importance of semantically codifying knowledge, and perceive semantic languages as a nuisance. Best practices are needed to help biologists create semantic ontologies [29], so that in the future a global and distributed group of high-quality RDF/OWL ontologies will be a reality.

The content of these ontologies should be non redundant and have common foundations, *e.g.* RO, for facilitating alignments and cross products.

Semantic languages, tools and interfaces Even though ontology editors, reasoners, APIs and Knowledge Base (KB) software for the Semantic Web have advanced a lot over the last few years, they still fall short of constituting established and robust technology, especially when it comes to their utility and reliability. On the language side, OWL is evolving fast and many new features are expected to appear in OWL 2[32].

The problems that we observed can only be addressed at the community level. Therefore, we make a public call for the creation and development of a Semantic System Biology community with the following aims:

1. Encourage and facilitate the creation of semantic bio-content.
2. Develop agreed upon best practices for such content creation.
3. Collect and index such content.
4. Agree upon and encourage a mechanism for identifying biological entities.
5. Facilitate the communication between the semantic technology developers and the life scientist, the users of such technology.

[29] http://bio2rdf.wiki.sourceforge.net/Banff+Manifesto
[30] http://lsrn.org/
[31] http://okkam.org/
[32] http://www.w3.org/TR/2008/WD-owl2-syntax-20080411/

This community should have objectives beyond those of the OBO Foundry: it should build upon the best of OBO (the community, the content creation guidelines, and the content) and exploit it in a standardized platform with emerging Semantic Web qualities. As a first step towards such a community, we are building the Semantic Systems Biology wiki[33].

We venture to consider the following topics to organize and structure the life sciences Semantic Web resourceome, and to define a set of principles:

1. orthogonality: avoid duplications of efforts
2. a defined set of RDF tags (*e.g.* definition, function, has_evidence, *etc.*)
3. a unique identifier per resource plus an ID resolution (*e.g.* purl.org)
4. comply to one agreed upon top-level ontology
5. comply to one agreed upon set of common relations
6. a list of prospective resource applications (*e.g.* hypothesis generation)
7. resource peer-review (community evaluation)
8. tooling (*e.g.* visualisation)
9. persistency-related issues
10. licensing (*e.g.* creative commons)
11. explicit semantics
12. rich axiomatisation (and hence rich querying)

We strongly feel this is the appropriate moment to establish such a community to bolster and extend the current efforts (*e.g.* HCLS IG, NeuroCommons[34]) and to begin building a universal, interoperable knowledge architecture [30]. Such a structured resource will further ensure that semantic technologies will become one of the most crucial means for knowledge integration in the life sciences [31].

Acknowledgements

This work was funded by the EU FP6 (LSHG-CT-2004-512143). EA was funded by the European Science Foundation (ESF) for the activity entitled Frontiers of Functional Genomics, ME by the University of Manchester and the EPSRC.

References

1. Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, S.M., Ogbuji, C., Rees, J., Stephens, S., Wong, G., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the semantic web. BMC Bioinformatics **8**(Suppl 3) (2007)
2. Clark, T., Kinoshita, J.: Alzforum and swan: the present and future of scientific web communities. Briefings in Bioinformatics **8**(3) (2007) 163–171

[33] http://www.bio.ntnu.no/systemsbiology/ssbwiki
[34] http://neurocommons.org

3. Villanueva-Rosales, N., Dumontier, M.: yowl: An ontology-driven knowledge base for yeast biologists. Journal of biomedical informatics (2008)
4. Belleau, F., Nolin, M.A.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics (2008)
5. Deus, H.F., Stanislaus, R., Veiga, D.F., Behrens, C., Wistuba, I.I., Minna, J.D., Garner, H.R., Swisher, S.G., Roth, J.A., Correa, A.M., Broom, B., Coombes, K., Chang, A., Vogel, L.H., Almeida, J.S.: A semantic web management model for integrative biomedical informatics. PLoS ONE **3**(8) (2008) e2946
6. Luciano, J.S.: Pax of mind for pathway researchers. Drug Discov Today **10**(13) (2005) 937–942
7. Post, L.J., Roos, M., Marshall, S.M., Driel, R., Breit, T.M.: A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data. Bioinformatics **23**(22) (2007) 3080–3087
8. Good, B.M., Wilkinson, M.D.: The life sciences semantic web is full of creeps! Brief Bioinform **7**(3) (2006) 275–286
9. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech **25**(11) (2007) 1251–1255
10. 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 Research **32**(Database-Issue) (2004) 262–266
11. Wheeler, D.L.L., Barrett, T., Benson, D.A.A., Bryant, S.H.H., Canese, K., Chetvernin, V., Church, D.M.M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y.Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J.J., Madden, T.L.L., Maglott, D.R.R., Miller, V., Ostell, J., Pruitt, K.D.D., Schuler, G.D.D., Shumway, M., Sequeira, E., Sherry, S.T.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L.L., Tatusova, T.A.A., Wagner, L., Yaschenko, E.: Database resources of the national center for biotechnology information. Nucleic Acids Res (2008)
12. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., Redaschi, N., su L. Yeh, L.: Uniprot: the universal protein knowledgebase. Nucleic Acids Res **32** (2004) 115–119
13. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., Rosse, C.: Relations in Biomedical Ontologies. Genome Biology **6** (2005) R46
14. Schulz, S., Kumar, A., Bittner, T.: Biomedical ontologies: what part-of is and isn't. J Biomed Inform **39**(3) (2006) 350–361
15. Antezana, E., Egaña, M., Baets, B.D., Kuiper, M., Mironov, V.: Onto-perl: An api for supporting the development and analysis of bio-ontologies. Bioinformatics **24**(6) (2008) 885–887
16. Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.A., Bogue, M., Booth, T., Brazma, A., Brinkman, R.R., Michael, Deutsch, E.W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J.M., Hardy, N.W., Hermjakob, H., Julian, R.K., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Novere, N.L., Leebens-Mack, J., Lewis,

S.E., Lord, P., Mallon, A.M., Marthandan, N., Masuya, H., Mcnally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J.M., Robertson, D.G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R.H., Schober, D., Smith, B., Snape, J., Stoeckert, C.J., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J., Wiemann, S.: Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. Nat Biotech **26**(8) (2008) 889–896
17. Quackenbush, J.: Standardizing the standards. Molecular Systems Biology **2** (2006)
18. Lam, H.Y., Marenco, L., Clark, T., Gao, Y., Kinoshita, J., Shepherd, G., Miller, P., Wu, E., Wong, G.T., Liu, N., Crasto, C., Morse, T., Stephens, S., Cheung, K.H.: Alzpharm: integration of neurodegeneration data using rdf. BMC Bioinformatics **8 Suppl 3** (2007)
19. Gao, Y., Kinoshita, J., Wu, E., Miller, E., Lee, R., Seaborne, A., Cayzer, S., Clark, T.: Swan: A distributed knowledge infrastructure for alzheimer disease research. J. Web Sem. **4**(3) (2006) 222–228
20. Cheung, K.H., Yip, K.Y., Smith, A., Deknikker, R., Masiar, A., Gerstein, M.: Yeasthub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics **21 Suppl 1** (2005) i85–i96
21. Neumann, E.K., Quan, D.: Biodash: A semantic web dashboard for drug development. In Altman, R.B., Murray, T., Klein, T.E., Dunker, A.K., Hunter, L., eds.: Pacific Symposium on Biocomputing, World Scientific (2006) 176–187
22. Smith, A.K., Cheung, K.H., Yip, K.Y., Schultz, M., Gerstein, M.K.: Linkhub: a semantic web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics **8 Suppl 3** (2007)
23. Pasquier, C.: Biological data integration using semantic web technologies. Biochimie (2008)
24. Lemoine, F., Labedan, B., Froidevaux, C.: Genoquery: a new querying module for functional annotation in a genomic warehouse. Bioinformatics **24**(13) (2008) i322–329
25. Laibe, C., Le Novere, N.: Miriam resources: tools to generate and resolve robust cross-references in systems biology. BMC Systems Biology **1**(1) (2007)
26. Mikel Egaña Aranguren, Wroe, C., Goble, C., Stevens, R.: In situ migration of handcrafted ontologies to reason-able forms. Data and Knowledge Engineering **66**(1) (2008) 147–162
27. Golbreich, C., Horrocks, I.: The OBO to OWL Mapping, GO to OWL 1.1! In: OWLED. (2007)
28. Mikel Egaña Aranguren, Bechhofer, S., Lord, P., Sattler, U., Stevens, R.: Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics **8** (2007) 57
29. Aranguren, M.E., Antezana, E., Kuiper, M., Stevens, R.: Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. BMC bioinformatics **9(Suppl 5)** (2008) S1
30. Slater, T., Bouton, C., Huang, E.S.: Beyond data integration. Drug Discovery Today **13**(13-14) (2008) 584–589
31. Sagotsky, J.A., Zhang, L., Wang, Z., Martin, S., Deisboeck, T.S.: Life sciences and the web: a new era for collaboration. Mol Syst Biol **4** (2008)