

# The RDF Foundry: Call for an initiative to build enhanced RDF resources for biological data integration

Aravind Venkatesan

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[venkates@nt.ntnu.no](mailto:venkates@nt.ntnu.no)

Ward Blondé

Ghent University,  
Department of Applied Mathematics,  
Biometrics and Process Control,  
Coupure links 653, 9000 Ghent,  
Belgium.

[ward.blonde@ugent.be](mailto:ward.blonde@ugent.be)

Erick Antezana

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[erick.antezana@bio.ntnu.no](mailto:erick.antezana@bio.ntnu.no)

Mats Skillingstad

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[skilling@alumni.ntnu.no](mailto:skilling@alumni.ntnu.no)

M Scott Marshall

Informatics Institute,  
University of Amsterdam,  
The Netherlands

[marshall@science.uva.nl](mailto:marshall@science.uva.nl)

Bernard De Baets

Ghent University,  
Department of Applied Mathematics,  
Biometrics and Process Control,  
Coupure links 653, 9000 Ghent,  
Belgium.

[bernard.debaets@ugent.be](mailto:bernard.debaets@ugent.be)

Vladimir Mironov

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[mironov@nt.ntnu.no](mailto:mironov@nt.ntnu.no)

Martin Kuiper<sup>\*</sup>

Norwegian University of Science and Technology (NTNU),  
Department of Biology,  
Høgskoleringen 5, 7491 Trondheim,  
Norway.

[kuiper@nt.ntnu.no](mailto:kuiper@nt.ntnu.no)

## ABSTRACT

Currently, the OBO Foundry plays an important role by setting guidelines to formalise the concepts within the biomedical domain. The ontologies within the OBO Foundry are usually represented in the OBO ontology language. While being human-readable, this language lacks the computational rigour required for the Semantic Web (SW). Consequently, the RDF and OWL languages, both fundamental components of the SW technology stack, are being increasingly adopted by the biomedical community to exchange biological knowledge in a computer intelligible form. Some of the OBO- formatted ontologies have been made available in OWL, thus signalling a move towards the

SW. OWL provides support for automated reasoning, which is its *raison d'être*. Unfortunately, automated reasoning on the massive volumes of data that are typical of the biomedical domain is riddled with performance limitations. Due to consistent support for the SPARQL specification in triple store implementations, as well as the ability to simulate some types of reasoning with pre-computed closures, RDF has evolved into a language of choice for knowledge exchange within the framework of the SW. Here, we discuss the need to establish a foundry charged with the task of harmonizing biomedical RDF resources, acting along the same lines as the OBO Foundry. To substantiate the need for an RDF Foundry, we provide the outcome of a small survey we have conducted to highlight the domain coverage, redundancies, and comprehensiveness of results obtained from a few representative distributed resources available today.

## General Terms

Management, Documentation, Performance, Design, Reliability, Experimentation, Standardization, Languages.

## Keywords

RDF, OWL, Knowledge Base, Data Integration, SPARQL, Query Federation, RDF Foundry.

---

<sup>\*</sup>Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WTMS'11, May 25–27, 2011, Sogndal, Norway.

Copyright 2011 ACM 978-1-4503-0148-0/11/05... \$10.00.

## 1. INTRODUCTION

The goal of the SW [1] is to make the World Wide Web computer intelligible. The SW languages OWL [2] and RDF [3] have become generally accepted by the SW community and by now they have made numerous forays into various application domains such as the Life Sciences [4].

RDF is currently the most widely adopted SW knowledge representation language. With properly designed RDF graphs the querying is very robust [5] and ultimately it is even possible to query multiple RDF stores. The SW Education and Outreach (SWE) Interest Group of World Wide Web Consortium (W3C) have initiated the Linking Open Data (LOD) [6] project for extending the present web content towards a RDF representation. In particular, the Linking Open Drug Data (LODD) [7] and the BioRDF [8] are among the several task forces in the Health Care and Life Sciences Interest Group (HCLS IG) [9] chartered by the W3C to promote the use of RDF in the life science and health care domain. As a consequence, a number of KBs [10, 11, 12, 13, 14] have mushroomed demonstrating the potential of SW technologies in life sciences and creating working applications that make direct use of ontological knowledge such as that supplied by the OBO Foundry.

The Vocabulary of Interlinked Datasets (VoID) [15, 29] is an emerging standard to facilitate the linking of various datasets by providing a common vocabulary to describe data in RDF Schema. Furthermore, the Banff Manifesto [16] provides some best practices for the design and implementation of RDF documents in the biological domain, for instance the use of normalized and de-referenceable URIs and mandatory predicates such as *rdfs:label* and *rdf:type*. Recent initiatives include the 3<sup>rd</sup> BioHackathon where those aspects have been raised once again [17].

Several ongoing activities within the HCLS IG are directed at providing guidelines for best practices in RDF as well as for federated querying [18]. The Scientific Discourse task force has worked on several ontologies that are meant to describe the origins of assertions that come from literature and experimental data, including SWAN/SIOC[19] and the Annotation Ontology[20]. A web based tool *aTags*[21], which uses a SIOC[22] for representing assertions in a consistent form of RDF has been used in HCLS IG for several projects. Concept Web Alliance (CWA) members based an initial proposal [23] for aggregating and sharing scientific statements on work in HCLS IG and W3C. Both the BioRDF task force and LODD task force have plans to produce W3C notes (i.e. specifications) in the coming months as guidelines for producing RDF specific to gene expression data (e.g. microarrays)[24], as well as for federation and linked data in general. The HCLS IG task forces are collaborating with others in the community and stakeholders, requesting feedback on the best practices under consideration.

The currently available SW Biological Knowledge Bases (KBs) [4] have certainly helped to demonstrate the advantages of the SW technologies, such as proper knowledge management, streamlined data integration and efficient SPARQL querying. However, the full potential of semantically encoded knowledge for querying, hypothesis generation and automated reasoning has not yet been realized. All the SW KBs constructed so far are essentially warehouses with all the classical shortcomings such as large up-front time investment required for data integration and querying, technical challenges with respect to the infrastructure,

data maintenance issues, data redundancy and a possible semantic mismatch of triples between various triple stores [18].

A plausible alternative to warehousing (built in the SW technologies) is to make use of federated queries over multiple triple stores from a single RDF endpoint [18, 25]. Recently an important advance has been made in this direction with the development of the SWObjects tool [30]. However, query federation is still in its nascent stage and has been hampered by the differences among KBs in the way they use RDF.

To establish the SW as a robust technology that facilitates the answering of complex biological questions, the above problems have to be addressed at the community level. Hereby we advocate the formation of a ‘foundry’ with the task of harmonizing the development of SW KBs in the biomedical domain that will work under the auspices of the HCLS IG. The RDF foundry will promote the interaction between the SW developers and biological data providers, encourage orthogonality of the content among the triple stores, and promote best practices for the use of RDF.

In the subsequent sections of this paper, we briefly review the major triple stores currently available for the biomedical domain, highlighting the redundancies in the content and querying hurdles.

## 2. Biological Knowledge Bases

The KBs chosen for our survey are listed in Table 1.

**Table 1: An overview of the surveyed KBs**

| KB               | No. of Triples        | No. of resources | No. of Endpoints | Triple Store Engine |
|------------------|-----------------------|------------------|------------------|---------------------|
| BioGateway       | ~1.8 billion triples  | 4                | 1                | Virtuoso            |
| Bio2RDF          | ~2.5 trillion triples | 40               | 40               | Virtuoso            |
| Linked Life Data | ~4.1 trillion triples | 22               | 1                | OWLIM               |
| HCLS KB – DERI   | ~400 million triples  | 15               | 1                | Virtuoso            |

BioGateway [11], hosted by the Norwegian University of Science and Technology (NTNU), integrates the entire set of OBO Foundry [26] ontologies (accepted and candidates), complete collection of annotations from the Gene Ontology Annotation (GOA) [27] files, fragments of NCBI taxonomy and SWISS-PROT. BioGateway uses two in-house developed ontologies (BioMetarel and MetaOnto) that provide a scaffold for data integration and semantic enrichment. The resources integrated in BioGateway share a common URI of the form <http://www.semantic-systems-biology.org> with each of the imported data sources represented with its individual graph name suffixed to the common URI. The store is augmented with pre-computed closures that increase the utility of RDF representation (see the accompanying paper Antezana et al). The SPARQL

query interface of BioGateway includes a large set of sample queries (both biology and ontology centric) which provide a starting point for the novices.

Bio2RDF [10] is a SW application developed and maintained by the Quebec Genomics Centre, Canada that provides a mashup of data from Gene Ontology, OMIM, Reactome, ChEBI BioCyc KEGG. Bio2RDF provides normalised URIs of the integrated resources of the form *http://bio2rdf.org/<namespace>:<identifier>* and unlike the other three KBs, Bio2RDF provides distributed endpoints corresponding to the individual resources instead of integrating the data in a single triple store. The URL for the various SPARQL endpoints is of the form *http://NAMESPACE.bio2rdf.org/sparql*.

Linked Life Data [13] is a semantic data integration platform developed by Ontotext as part of the Large Knowledge Collider (LarKC) project. The platform interconnects datasets from the Pathway and Interaction KB (PIKB) [27], PubMed, KEGG, IntAct, MINT, Entrez-Gene, and SKOS representation of OBO ontologies. The integrated resource URIs in Linked Life Data are of the form *lld:resource/db/type/id*. As a convention, this KB retains the original RDF structure if distributed by the data provider and uses resolvable URIs for data sources with no RDF distributions.

The HCLS KB is hosted at DERI, Galway [28]. The contents of this KB are mainly derived from two sources: the Neurocommons KB [14] and the LODD. The Neurocommons KB provides RDF/XML versions of resources which include OBO (including the Gene Ontology), MEDLINE, Gene Ontology Annotation (GOA), Medical Subject Headings (MeSH), and parts of

**Table 2: The querying results of CDC2**

| Knowledge Base   | Triple Store Endpoint                                                                                                                 | No. of Triples retrieved by Q1 | No. of property types associated with | No. of Triples retrieved by Q2 | No. of triples with <i>rdfs:seeAlso</i> property |
|------------------|---------------------------------------------------------------------------------------------------------------------------------------|--------------------------------|---------------------------------------|--------------------------------|--------------------------------------------------|
| BioGateway       | <a href="http://www.semantic-systems-biology.org/biogateway/querying">http://www.semantic-systems-biology.org/biogateway/querying</a> | 146                            | 20                                    | 50                             | 0                                                |
| Bio2RDF          | <a href="http://uniprot.bio2rdf.org/sparql">http://uniprot.bio2rdf.org/sparql</a>                                                     | 265                            | 28                                    | 0                              | 56                                               |
| Linked Life Data | <a href="http://linkedlifedata.com/sparql">http://linkedlifedata.com/sparql</a>                                                       | 354                            | 27                                    | 0                              | 68                                               |
| HCLS KB - DERI   | <a href="http://hcls.deri.org/sparql">http://hcls.deri.org/sparql</a>                                                                 | 18                             | 6                                     | 15                             | 0                                                |

SenseLab neurobiology databases. The datasets generated by LODD includes DailyMed, DrugBank, Diseasesome and SIDER to name a few. The resources in this KB are divided into named graphs and the URIs is of the form *http://hcls.deri.org/resource/graph/graphName*.

### 3. Querying

First, we used a generic query (Q1 below) to retrieve the neighbourhood of a concept present in all the four KBs. For this survey, the extensively studied human CDC2 protein kinase (UniProt accession: P06493) was used.

Q1 (with the HCLS KB URL):

```
PREFIX term_id: <http://purl.org/obo/owl/IMR#IMR_0704386>
SELECT distinct ?outwardarrow ?head_id ?tail_id ?inwardarrow
WHERE {
  {
    term_id: ?outwardarrow ?head_id.
  } UNION {
    ?tail_id ?inwardarrow term_id:.
  }
}
```

As can be seen in Table 2, the query dutifully retrieves the expected triples from all the four KBs, showing a large difference in the volume of data associated with this protein in the surveyed KBs. However, the output can hardly be palatable for any human being because of the often inscrutable URIs (an example in Table 3). This begs the question: who are the intended users of the SW KBs being currently developed? Are they meant for life science researchers or for the SW developers themselves? If the former is the case (as it should be) then we need a different query.

The query Q2 below is expected to produce a human-readable output. As Q1, it shows the neighbourhood of CDC2 but this time with the connector *rdfs:label*.

Q2 (with the HCLS KB URL):

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

```
PREFIX term_id: <http://purl.org/obo/owl/IMR#IMR_0704386>
SELECT distinct ?subject ?predicate ?object
WHERE {
  {
    term_id: ?outwardarrow ?head_id.
    term_id: rdfs:label ?subject.
    ?outwardarrow rdfs:label ?predicate.
    ?head_id rdfs:label ?object.
  } UNION {
```

```

?tail_id ?inwardarrow term_id.
?tail_id rdfs:label ?subject.
?inwardarrow rdfs:label ?predicate.
term_id: rdfs:label ?object.
}
}

```

This query returns the expected results from BioGateway and HCLS KB (Table 4) but not from Bio2RDF or Linked Life Data. After investigation, we found that *rdfs:label* is not used in the latter two KBs in the way we have anticipated, so our query does not match any labels. In general, there is a striking difference among the surveyed KBs in the way they use properties to describe the same concept (Table 2). In particular, we observed an extensive use of the property *rdfs:seeAlso* in some cases. This property is essentially equivalent to the 'hyperlink' and its use has no well-defined meaning or semantics.

**Table 3: Examples of Q1 results**

| outwardarrow | head_id                                                                                             |
|--------------|-----------------------------------------------------------------------------------------------------|
| rdf:type     | uniprot:Protein                                                                                     |
| rdfs:seeAlso | <a href="http://purl.uniprot.org/interpro/IPR011009">http://purl.uniprot.org/interpro/IPR011009</a> |
| rdfs:seeAlso | <a href="http://purl.uniprot.org/interpro/IPR000719">http://purl.uniprot.org/interpro/IPR000719</a> |
| rdfs:seeAlso | <a href="http://purl.uniprot.org/pfam/PF00069">http://purl.uniprot.org/pfam/PF00069</a>             |

## 4. The RDF Foundry

### 4.1 Mission Statement

RDF and its query language SPARQL are excellent technologies for integrating and using large amounts of biomedical knowledge. However, we find RDF too unstructured to be used productively without any further conventions. Therefore, we propose some guidelines on RDF representations that are meant to be integrated with other representations in a large store and used in combination with other stores. These guidelines will enable generic SPARQL queries instead of data-specific queries and consequently federated querying. We encourage everybody to engage in a broad discussion of the proposed guidelines with the purpose of building consensus around further refinement.

The RDF Foundry will:

- promote orthogonality among resources,
- identify important gaps in the resources,
- establish and maintain a single top-level ontology,
- maintain a single unique identifier per resource,
- enable distributed queries among resources,
- facilitate the integration of results,

**Table 4: Examples of Q2 results retrieved from BioGateway and HCLS KB**

| Knowledge Base | Subject    | Predicate         | Object                    |
|----------------|------------|-------------------|---------------------------|
| BioGateway     | CDC2       | has function      | protein binding           |
|                | CDC2       | is located in     | spindle microtubule       |
|                | CDC2       | has source        | Homo sapiens              |
|                | CDC2       | interacts with    | CDC25C                    |
| HCLS KB - DERI | CDC2_HUMAN | type              | Class                     |
|                | CDC2_HUMAN | has_dbxref        | UniProt:P06493            |
|                | CDC2_HUMAN | has_exact_synonym | CDC2                      |
|                | CDC2_HUMAN | has_exact_synonym | Cyclin-dependent kinase 1 |

- promote exchange of resources.

### 4.2 Proposed guidelines

RDFS was designed to provide a consistent structure for RDF. Many of its ideas are widely used, like the connector *rdfs:label* between a resource identifier and its human-readable name. One particular feature of RDFS is that it does not make a distinction between instances and classes, which is not compatible with the current ontology practices. The RDF Foundry will promote a universal distinction between instances and classes. This practice will sustain the development and use of ontologies.

The design of OWL is another example of imposing restrictions on modelling in RDF. Very early in the history of its design, OWL was split in three different fragments: OWL Full, OWL DL and OWL Lite. For the sake of backward compatibility with RDFS, OWL Full does not make a distinction between instances and classes either. This reduces the utility of OWL Full and RDFS with respect to ontologies essentially to informal modelling practices. In contrast, OWL DL and OWL Lite are based on description logics, which imply a strict distinction between instances and classes. Other description logic based fragments were designed later on for OWL, like OWL 2 EL, OWL 2 RL and OWL 2 QL. Unfortunately, all these fragments make use of the so-called *defined classes*, often represented by blank nodes in RDF. The defined classes, while having a proper identifier, are often poorly (or even not at all) named and typically lack a definition in natural language. Their logic based definitions might be very useful for consistency checking and classification during the design phase of an ontology, but they turn into real hurdles for the domain expert when it comes to knowledge exploration (e.g. by browsing or querying) and hypothesis generation. The defined classes call for unintuitive design practices which are difficult for the end-users to follow and if used for automated reasoning they are prone to generating inconsistencies. The RDF Foundry will insist on every class being represented as a primitive (user defined) class with a human readable name.

In the RDF Foundry, relations between classes should be represented as triples of the form **subject-predicate-object**,

where the predicate represents only a necessary condition between two primitive classes. This will enable the use of an important reasoning technique called logic entailment. The logical inference of implicit knowledge can be implemented by closure rules like reflexivity, transitivity, priority over subsumption, super-relations and chains of relations. The relation types used should be assembled and maintained in a separate RDF graph, central in the RDF Foundry that uses Metarel or OWL vocabulary for the annotation of their semantics. By avoiding logically defined classes, the RDF representations will not contain sets of relation triples that form sets of necessary and sufficient conditions for defining a class. **Every triple will be considered as a knowledge statement that holds in itself and for which the interpretation is independent of the rest of the RDF graph.** This is a central idea that makes RDF intelligible for scientists who are not familiar with logics.

Well-modelled RDF needs additional guidelines other than a universal distinction between instances and classes and the independence of the interpretation of RDF triples. Some of our suggestions are listed below and we invite other suggestions from all the interested parties.

- Standard practices for a uniform way of providing names, definitions, synonyms, cross-references, etc.
- Giving RDF graphs URI's and then making statements about those URI's in VoiD and other metadata vocabularies.
- Complete data provenance [24].

## 5. ACKNOWLEDGMENTS

This work is originally funded by the Norwegian University of Science and Technology (NTNU), Norway. WB was funded by Blondé Engineering. VM was funded by FUGE Mid-Norway.

## 6. REFERENCES

- [1] Berners-Lee T and Hendler J. 2001. Publishing on the Semantic Web. *Nature*. 410:1023-1024.
- [2] Resource Description Framework (RDF) Model and Syntax Specification, Nov 2010. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>
- [3] <http://www.w3.org/TR/owl-ref/>
- [4] Antezana E. et al. 2009. Biological knowledge management: the emerging role of SW technologies. *Brief Bioinform* 10:392-407.
- [5] Mironov V et al. 2010. Benchmarking triple stores with biological data. *Semantic Web Applications and Tools for Life Sciences (SWAT4LS)*, Berlin, Germany. December 10, 2010.
- [6] <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>
- [7] <http://www.w3.org/wiki/HCLSIG/LODD>
- [8] [http://www.w3.org/wiki/HCLSIG\\_BioRDF\\_Subgroup](http://www.w3.org/wiki/HCLSIG_BioRDF_Subgroup)
- [9] HCLS IG: <http://www.w3.org/blog/hcls>
- [10] Belleau F et al. 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform.* 41(5): 706-16.
- [11] Antezana E et al. 2009. BioGateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics*. 1:10 Suppl 10:S11.
- [12] Antezana E et al. 2009. The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process, *Genome Biology*, 10: R58
- [13] Linked Life Data: <http://www.linkedlifedata.com/>
- [14] Ruttenberg A et al. (2009). Life sciences on the SW: the Neurocommons and beyond. *Brief Bioinform.* 10 (2):193-204.
- [15] VoiD: <http://semanticweb.org/wiki/VoiD>
- [16] Banff Manifesto, HCLS DI, Canada, 2007. [http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff\\_Manifesto](http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto)
- [17] Third BioHackaton, Tokyo, Japan; 2010. [http://hackathon3.dbcls.jp/wiki/Data\\_Provider\\_Session](http://hackathon3.dbcls.jp/wiki/Data_Provider_Session), <http://hackathon3.dbcls.jp/wiki/URI>
- [18] Cheung KH et al. 2009. A journey to Semantic Web query federation in the life sciences., *BMC bioinformatics*, 10 Suppl 10, S10.
- [19] SWAN/SIOC: Alignment Between the SWAN and SIOC Ontologies, <http://www.w3.org/TR/hcls-swansioc/>, W3C Note 2009
- [20] AO: <http://code.google.com/p/annotation-ontology/wiki/Homepage>
- [21] Samwald and Holger Stenzhorn. 2010. "Establishing a distributed system for the simple representation and integration of diverse scientific assertions." *Journal of Biomedical Semantics* 1, Nr. 1: S5. (journal article)
- [22] SIOC, SIOC Types and Health Care and Life Sciences, <http://www.w3.org/TR/hcls-sioc/>, W3C Note 2009
- [23] The anatomy of a nanopublication, by: Paul Groth, Andrew Gibson, Jan Velterop. *Information Services and Use*, Vol. 30, No. 1. (1 January 2010), pp. 51-56. doi:10.3233/ISU-2010-0613 Key: citeulike:7927702
- [24] Deus HF et al. 2010. Provenance of Microarray Experiments for a Better Understanding of Experiment Results. *Proceeding of the second International Workshop on the role of Semantic Web in Provenance Management (SWPM2010, ISWC)*, Shanghai, China.
- [25] Cheung KH et al. 2009. Semantic Web for Health Care and Life Sciences: a review of the state of the art., *Briefings in Bioinformatics*, 10 (2), 111-3
- [26] Smith B et al. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology*, 25(11), 1251-1255.
- [27] PIKB: <http://www.ontotext.com/lifeskim/pikb.html>.
- [28] [http://www.w3.org/wiki/HCLSIG\\_BioRDF\\_Subgroup/DERI\\_HCLS\\_KB](http://www.w3.org/wiki/HCLSIG_BioRDF_Subgroup/DERI_HCLS_KB).
- [29] Describing Linked Datasets with the voiD Vocabulary: <http://www.w3.org/2001/sw/interest/void/>.
- [30] Prud'hommeaux E et al. Tutorial: Query Federation with SWObjects. Available from *Nature Precedings*