Facilitating integrated analysis of biological data by enhancing interoperability of RDF resources: Practical Recommendations.

Aravind Venkatesan[a*], Ward Blondé[b], Erick Antezana[a], M Scott Marshall[c], Andrea Splendiani[d], Mikel Egaña Aranguren[e], James Malone[f], Vladimir Mironov[a] and Martin Kuiper[a*]

[a]*Norwegian University of Science and Technology (NTNU), Department of Biology, Høgskoleringen 5, 7491 Trondheim, Norway. Email: {venkates, mironov, kuiper} @nt.ntnu.no, erick.antezana@bio.ntnu.no.*
[b]*Institute for Medical Informatics, Statistics and Documentation, LKH-Eingangsgebäude, Auenbruggerplatz 2, 8036 Graz, Austria. Email: ward.blonde@medunigraz.at.*
[c] *Department of Medical Statistics and Bioinformatics, Leiden University Medical Center / Informatics Institute, University of Amsterdam. Email: marshall@science.uva.nl.*
[d]*Biomathematics and Bioinformatics Department, Rothamsted Research, West Common, Harpenden, Hertfordshire, AL5 2JQ. Email: andrea.splendiani@rothamsted.ac.uk*
[e]*School of Computer Science, Technical University of Madrid (UPM), Spain. Email: megana@fi.upm.es*
[f]*European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom. Email: malone@ebi.ac.uk.*

Abstract. The Resource Description Framework (RDF) has evolved into a language of choice for knowledge exchange on the (Semantic) Web due to its robustness and queribility with SPARQL. However, RDF allows for a variety of modelling practices which, if used unchecked, would result in the production of resources that may be only partially compatible with other resources. We highlight the issues related to incomplete interoperability that exist today with the results of a limited survey of the current interoperability of public RDF resources. We propose a set of informed guidelines that can help in overcoming the problem.

Keywords: Semantic Web, RDF, RDF Recommendations, RDF resources, SPARQL, Data Integration, Knowledge Base, Query Federation.

1. Introduction

The Life Science community is challenged by a data deluge owing to the development of high-throughput technologies. Although these technologies have revolutionised and transformed the research in Life Sciences, efficient integration of the generated data remains a demanding task, since this requires an effective representation of the underlying knowledge: a representation that is structured, reusable and machine-friendly. The Semantic Web (SW) Technology [8] essentially aims at transforming the current World Wide Web (WWW) into a global reasoning and semantics-driven Knowledge Base (KB). Languages such as the Resource Description Framework (RDF) [20] and the Web Ontology Language (OWL) [9] have been developed to support this endeavor. Currently, we are witnessing a growing acceptance of SW technologies in the Life Science domain as a means to manage biological knowledge.

Unlike some other scientific domains where knowledge is largely based on mathematical representation, the biomedical domain relies to a large

[*] Corresponding author(s)

extent on natural language descriptions and inferences based on the understanding of the characterised entities [32]. For example, biological sequence similarities are determined by using mathematical algorithms, which are used in turn by biomedical scientists for annotations about biological functions. Bio-ontologies were developed with a vision of capturing knowledge and characterising entities in the biological domain. These are widely used by the biomedical scientists to make inferences about the uncharacterised entities.

Most bio-ontologies are developed in the OBO format, originally introduced by the Gene Ontology initiative [5], which is a format primarily meant for human comprehension. The OBO Foundry [32] was established to coordinate the formalisation of the concepts of the biomedical domain according to clear and sound principles, which include the commitment to use shared relations and potentially a common upper-ontology perspective [3]. Although some biomedical ontologies were developed primarily in OWL (for instance, the Protein Ontology [23]) or provide an OWL version of their ontology such as the Cell Cycle Ontology [2], it has been observed that there are so far very few successful implementations that exploit the full capabilities of automated reasoning offered by OWL. The “less expressive” language RDF, on the other hand, has enabled handling of large amounts of knowledge due to its simplicity. The graph-based data model of RDF makes it a compelling choice to model knowledge and integrate data from multiple sources. It has become the cornerstone for data integration across computing platforms due to its flexibility and its suitability to represent concepts in the biomedical domain. Through the query language SPARQL [26], users are provided with the capability of simultaneously querying and integrating results from multiple RDF graphs. With properly designed RDF graphs the querying is very robust [36] and in principle it is even possible to query multiple RDF stores.

In the subsequent sections of this paper we briefly review the current initiatives taken to improve linking of various datasets in RDF. In addition, we present the results of a survey of major triple stores available for the biomedical domain, in which we focus on interoperability issues and highlight the querying hurdles from a user’s perspective. We finish with some suggestions on how best to represent knowledge so as to avoid the described problems.

2. Refining RDF representations

RDF is currently the most widely adopted SW knowledge representation language. However, along with all the advantages of RDF outlined in the previous section, some limitations originate from RDF’s major strength: its flexibility. RDF allows for a variety of modelling practices, which, if used unchecked, may result in the production of incompatible resources. Therefore, a number of initiatives have been undertaken to harmonise the use of RDF. Most importantly, the SW Education and Outreach Interest Group (SWEOIG) [33] of the World Wide Web Consortium (W3C) has initiated the Linking Open Data (LOD) project [21] for extending the present web content towards an RDF representation. A number of projects have been initiated for the effective usage of RDF. The LOD project provides a list of recommendations including the use of 'cool' URIs [31] and the consistent use of descriptive predicates like *rdfs:label*. The Vocabulary of Interlinked Datasets (VoID) [4] is an emerging standard to facilitate the linking of various datasets by providing a common vocabulary to describe data in RDF Schema. The Banff Manifesto [18] provides some best practices for the design and implementation of RDF documents in the biological domain, including a) the use of normalised and dereferencible URIs; b) mandatory predicates such as *rdfs:label* and *dc:title*; c) prohibiting the use of blank nodes. Furthermore, the Semantic Web Health Care and Life Sciences Interest Group (HCLS IG) provides guidelines for best practices in RDF via several ongoing activities, such as the alignment of the Semantic Web Applications in Neuromedicine (SWAN, an ontology for the description of scientific discourse), and the Semantically Interlinked Online Communities (SIOC, ontology to integrate online community information) ([13], [24], [25]). This offers a model to make scientific discourses in online communities computationally more viable. The web based tool *aTags* [30], which uses SIOC for representing assertions in a consistent form in RDF, has been used by HCLS IG for several projects [11]. The Concept Web Alliance (CWA) developed an initial proposal of the nano-publication model that enables the aggregation of fine-grained scientific information across the web in RDF [17], and, last in our non-exhaustive overview, members of the BioRDF and LODD task forces have produced substantial literature on guidelines of how to produce RDF [15, 16, 34].

3. Biological Knowledge Bases

The growing importance of RDF as a standard has encouraged the adoption of SW technologies in the Life Sciences [1, 6, 29, 35]. This has also encouraged data providers such as UniProt [28] and the Gene Expression Atlas [19] to publish their data in RDF. The currently available SW Biological Knowledge Bases (KBs) have certainly helped to demonstrate the advantages of the SW technologies, including a richer knowledge representation, streamlined data integration and efficient SPARQL querying. This section provides a brief description of these KBs.

Bio2RDF [6] is a SW application developed and maintained by the Quebec Genomics Centre, Canada, that provides a mashup of data from the likes of the Gene Ontology, OMIM, Reactome, ChEBI, BioCyc and KEGG. Bio2RDF provides normalised URIs of the integrated resources of the form *http://bio2rdf.org/<namespace>:<identifier>* and unlike the other three KBs described in this section, 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*.

The **Neurocommons** [29] project aims at supporting research for neurological diseases. The 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 the SenseLab neurobiology databases.

The **HCLS KB** is hosted at DERI, Galway [14]. The content of this KB is mainly derived from two sources: the Neurocommons KB and the Linking Open Drug Data [22]. LODD data includes DailyMed, DrugBank, Diseasome and SIDER, to name a few. The resources in this KB are divided into named graphs and the URIs are of the form *http://hcls.deri.org/resource/graph/graphName*.

BioGateway [1] integrates the entire set of OBO Foundry ontologies (including both accepted and candidate OBO ontologies), the complete collection of annotations from the Gene Ontology Annotation (GOA) files, and fragments of the NCBI taxonomy and SWISS-PROT. BioGateway also uses two relation ontologies (BioMetarel and MetaOnto) specifically developed to 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 prefixed with the common URI. The store is augmented with pre-computed closures that increase the utility of the RDF representation [10]. The SPARQL query interface of BioGateway includes a large set of sample queries (both biology and ontology-centric) that provide a starting point for the novices.

Linked Life Data (LLD) [35] 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), PubMed, KEGG, IntAct, MINT, Entrez-Gene, and the SKOS representation of OBO ontologies. The integrated resource URIs in LLD 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.

4. What is missing

Although the SW KBs provide a proof of concept, the full potential of semantically encoded knowledge for querying, hypothesis generation and automated reasoning has not yet been realised. All the SW KBs constructed so far are essentially warehouses with all the classical shortcomings such as a large up-front time investment required for data integration and querying, technical challenges with respect to the infrastructure, data provenance and maintenance issues, data redundancy and a possible semantic mismatch of triples between various triple stores. This being said, even at this stage SW technologies offer an interesting alternative to warehousing, as they offer a native support for explicit semantic definition and the possibility of federated queries being directed to multiple triple stores from a single RDF endpoint. Recently, an important advancement in this direction has been made with the release of SPARQL 1.1 and the development of promising projects such as the SWObjects tool [27]. However, query federation is still in its nascent stage and has been hampered by

Table 1: An Overview of the surveyed KBs

Knowledge Bases	No. of Triples	No. of resources	No. of Endpoints	Triple Store Engine
BioGateway	~1.8 billion	4	1	Virtuoso
Bio2RDF	~2.5 trillion	40	40	Virtuoso
Linked Life Data	~4.1 trillion	22	1	OWLIM
Neurocommons	~350 million	20	1	Virtuoso

the differences among KBs in the way they use RDF [11, 12].

The establishment of SW as a robust technology in the Life Science domain depends greatly on how this caters for the end-users' (biologists') needs. Further development and application of SW technologies should go in parallel with the adoption of the integrated resources by the users. Therefore it is important to understand to which extent current technologies are limited in their adoption by the maturity of their implementation. In the Life Sciences domain we observe that heterogeneous representations hamper the ability to effectively perform SPARQL queries from users. We demonstrate this through a simple test, based on which we propose an initial set of suggestions which can overcome the observed limitations. Specifically, we perform a few common application oriented queries to integrated RDF biomedical datasets, and inspect the results.

Table 1 lists the KBs that were considered for the queries. 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 NeuroCommons 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 it can be seen from Table 2, the query retrieves the expected triples from all four KBs, showing a large difference in the volume of data associated with this protein in the surveyed KBs. This query is quite valuable, because it makes the KBs browsable, as defined by Tim Berners-Lee in his proposal for Linked Data [7]. However, the output displays a list of URIs (an example is provided in Table 3), and a more elaborated query has to be developed to retrieve human readable labels.

Linked Data guidelines promote the use of at least one of the four properties to be used between the resource identifiers and their human-readable name, which includes *rdfs:label*, *foaf:name*, *skos:prefLabel* and *dcterms:title*. In order to produce a human-readable output the query Q2 (see below) was formulated. The KBs were queried using all the four properties mentioned above, the example below demonstrates the use of the *rdfs:label* predicate.

Q2 (with NeuroCommons 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.
  }
}
```

Table 2: The querying results for CDC2

	BioGateway	**Bio2RDF**	**Linked Life Data**	**Neurocommons**
Q1	265	146	5146	19
Q2 – *rdfs:label*	50	0	0	14
Q2 – *skos:prefLabel*	0	0	0	0
Q2 – *foaf:name*	0	0	0	0
Q2 – *dcterms:title*	0	0	0	0
Q3 – *rdfs:label*	50	1	0	14
Q3 – *skos:prefLabel*	0	0	0	0
Q3 – *foaf:name*	0	0	0	0
Q3 – *dcterms:title*	0	0	0	0

This query returns the expected results only from BioGateway and Neurocommons and only for *rdfs:label* (Tables 2 and 4). The absence of any output from Bio2RDF or LLD was particularly startling.

Therefore, we modified the query Q2 to produce Q3 in an attempt to identify the reason. In Q3 the properties are used only for the subject and the object but not for the predicate.

Table 3: Examples of Q1 results

Term_id	**Outwardarrow**	**Head_id**
http://bio2rdf.org/uniprot:P06493	http://purl.uniprot.org/core/citation	http://bio2rdf.org/citations:15489334
http://purl.uniprot.org/uniprot/P06493	rdfs:seeAlso	http://purl.uniprot.org/interpro/IPR011009
http://purl.org/obo/owl/IMR#IMR_0704386	http://www.geneontology.org/formats/oboInOWL#hasExactSynonym	nodeID://1002742944
http://www.semantic-systems-biology.org/ SSB#P06493	http://www.semantic-systems-biology.org/SSB#has_function	http://www.semantic-systems-biology.org/SSB#GO_0005515

Table 4: Examples of Q2 results retrieved from BioGateway and Neurocommons KBs

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
Neurocommons	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

Q3 (with NeuroCommons 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: ?predicate ?head_id.
  term_id: rdfs:label ?subject.
  ?head_id rdfs:label ?object.
 }
 UNION
 {
  ?tail_id ?predicate term_id:.
  ?tail_id rdfs:label ?subject.
  term_id: rdfs:label ?object.

 }
}
```

With this query format, human-readable output was returned with the use of *rdfs:label* from Bio2RDF, however the query did not return any results for LLD (Tables 2 and 5).

The queries Q2 and Q3 demonstrates that not all representation use *rdfs:label* (or other properties) with the same commitment. The above query could be refined further by using both UNION, to cope explicitly with each property, or OPTIONAL, to cope in a consistent way with missing values. However, this would make the above query, which is intrinsically simple, rather complex. Even worse, the author of the query would need to know all possible representation variants to compensate for the lack of homogeneity in the RDF sources.

As illustrated by our examples, there is clearly a considerable difference among the surveyed KBs in the way they use properties to describe the same concept (Table 2). In real Life Science use cases, a SW specialist would be required to spend considerable time to understand the layout of the KBs to formulate suitable queries (based on the biological question) in order to produce a consolidated result for the biologists.

In order to successfully exploit the possibilities for

Table 5: Example of Q3 results retrieved from Bio2RDF, Neurocommons and BioGateway

Knowledge Base	Subject	Predicate	Object
Bio2RDF	CDC2	http://www.w3.org/2002/07/owl#sameAs	CDC2_HUMAN [uniprot:P06493]
Neurocommons	CDC2_HUMAN	http://www.geneontology.org/formats/oboInOwl#hasDbXref	UniProt:P06493
BioGateway	CDC2	http://www.semantic-systems-biology.org/SSB#has_function	protein binding

query and query federation in SPARQL, differences between stores will have to be reduced as much as possible.

5. Suggestions for further alignment of KBs

RDF and its query language SPARQL are excellent technologies for integrating and using large amounts of biomedical knowledge. However, the aforementioned heterogeneity must be addressed to provide real value for end users, in our case Life Sciences researchers. RDF best practices to improve machine and human readability should be discussed and addressed at the community level. As an initial step, we propose some suggestions that are meant to be complementary to the practices followed by the main RDF Life Sciences providers. These suggestions will simplify the development of SPARQL queries that can consistently retrieve data from different resources, thus making effective use of query federation:

- devise best practices that enable distributed queries among resources;
- devise best practices that facilitate the integration of results;
- promote the exchange of resources;
- promote the consistent use of metadata;
- identify important gaps in the resources;
- provide complete documentation to enable users to write queries addressing representative use cases;
- investigate the practice of URI ‘normalisation’ whereby existing resources are renamed using a new URI (e.g. [6]);
- identify existing technologies to build tools that bring the technology closer to the user community.

5.1. Specific recommendations

First of all, a wider acceptance of the Linked Data recommendations should be promoted, including:

- The universal use of de-referencible HTTP URIs.
- Preferentially only the most basic (and commonly used) features of RDF should be used.
- Features like reification, collections, containers, and blank nodes should be best avoided.
- The use of descriptive properties (*rdfs:label*, *skos:prefLabel*, *foaf:name*, *dcterms:title*) should be consistent to enhance human readability.
- Extensive linking of the data.

Additionally we recommend to:

- Keep the set of instances and the set of classes disjoint in order to guarantee compatibility with decidable flavors of OWL, like OWL EL.

- Provide natural language definitions of all classes.

These recommendations are expected to greatly improve the interoperability of KBs, and would only require minor tuning of the existing semantic KBs**.** However, it is anticipated that data providers will comply with these recommendation to varying degrees. Therefore, it would be important to make users aware of the degree of compliance of the available semantic resources. This may be achieved through a wide adoption of a certification system reflecting that degree. To this end, a five-level certification scheme is proposed, inspired by the 5 level system introduced for Linked Open Data [7]. The scheme assumes a full compliance with the 5-star system proposed by W3C and is supposed to operate on top of it:

*	Valid RDF with de-referencible HTTP URIs.
**	Human-friendly labelling of resources (e.g. *rdfs:label*) and natural language definitions for classes.
***	At least a minimal set of metadata, first of all data description, provenance and availability.
****	No blank nodes, containers, collections or reification.
*****	A mirrored version in OWL EL for consistency checking and automated reasoning.

Each level assumes compliance with the lower levels. This system takes into account two factors: 1) the feasibility; and 2) the likelihood of acceptance by data providers and data consumers. From this point of view the first level is *sine qua non*. The ** and *** levels are both desirable and feasible, and are likely to be adopted by the community. The **** level is feasible in most cases though not all and may pose an additional burden on the developers. Finally the ***** level is associated with considerable investments for data providers and needs to be investigated further.

6. Conclusions

RDF and its query language SPARQL are excellent technologies for integrating and using large amounts of biomedical knowledge. However, to establish the RDF as a robust technology that facilitates hypothesis generations and answering of complex biological questions for a wide user community the issues raised above have to be addressed at the community level.

The HCLS IG has already provided an overview of emerging best practices in the usage of RDF [34]. We propose that the interested parties engage in a broad discussion of these recommendations and the certification scheme suggested above with the aim of building a community wide consensus. This goal could be achieved only through active interaction between SW developers, biological data providers and life scientists.

We believe the aforementioned suggestions for RDF representations will facilitate integration of biomedical resources and will enable more generic SPARQL queries instead of data-specific ones and consequently more efficient federated querying.

References

[1] E. Antezana, W. Blonde, M. Egana, A. Rutherford, R. Stevens, B. De Baets, V. Mironov, and M. Kuiper. BioGateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics,* 10 2009 Suppl 10**,** S11.

[2] E. Antezana, M. Egana, W. Blonde, A. Illarramendi, I. Bilbao, B. De Baets,, R. Stevens, V. Mironov, and M. Kuiper. The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. *Genome Biology,* 10 (2009)**,** R58.

[3] R. Arp and B. Smith. Function, Role, and Disposition in Basic Formal Ontology. Available from Nature Precedings *In:* Nature Proceedings, 2008.

[4] K. Alexander, R. Cyganiak, M. Hausenblas and J. Zhao 2010. Describing Linked Datasets with the VoID Vocabulary. W3C, available at http://www.w3.org/2001/sw/interest/void/

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G.M.Rubin, and G. Sherlock, Gene ontology: tool for the unification of biology. *Nature Genetics,* 25 (2000)**,** 25-9.

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

[7] T. Berners-Lee 2006. Linked Data, available at http://www.w3.org/DesignIssues/LinkedData.html.

[8] T. Berners-Lee and J. Hendler, Publishing on the semantic web. *Nature,* 410 (2001)**,** 1023-4.

[9] S. Bechhofer, V. F. Harmelen, J. Hendler, I. Horrocks, D. L. Mcguinness, P. F. Patel-Schneider, and L. A. Stein 2004. Web Ontology Language (OWL). W3C, available at http://www.w3.org/TR/owl-ref/ .

[10] W. Blonde, V. Mironov, A. Venkatesan, E. Antezana, B. De Baets, and M. Kuiper, Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinformatics,* 27 (2011)**,** 1562-8.

[11] K. H. Cheung, H. R. Frost, M. S. Marshall, E. Prud'hommeaux, M. Samwald, J. Zhao, and A. Paschke. A journey to Semantic Web query federation in the life sciences. *BMC Bioinformatics,* 10 (2009), Suppl 10**,** S10.

[12] K. H. Cheung, E. Prud'hommeaux, Y. Wang, and S. Stephens. Semantic Web for Health Care and Life

Sciences: a review of the state of the art. *Briefings in Bioinformatics,* 10 (2009)**,** 111-3.

[13] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg and T. Clark. The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics,* 41 (2008)**,** 739-51.

[14] DERI Health Care and Life Science Knowledge Base 2009. Available at http://www.w3.org/wiki/HCLSIG_BioRDF_Subgroup/DERI_HCLS_KB.

[15] H. Deus et al. (2010). Provenance of Microarray Experiments for a Better Understanding of Experiment Results. *In:* Proceeding of the second International Workshop on the role of Semantic Web in Provenance Management (SWPM2010, ISWC), 2010 Shanghai, China.

[16] H. Deus et al. Translating standards into practice – one semantic web API for gene expression. *Journal of Biomedical Informatics,* Available online 24 March 2012, ISSN 1532-0464.

[17] P. Groth, A. Gibson and J. Veleltrop The anatomy of a nanopublication. *Information Services and Use,* 30 (2010)**,** 51-56.

[18] HCLS-DI Banff Manifesto 2007, Banff. Available at http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto.

[19] M. Kapusheky, I. Emam, E. Holloway, P. Kurnosov, A. Zorin, J. Malone, G. Rustici, E. Williams, H. Parkinson, and A. Brazma. Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research,* 38 (2010)**,** D690-D698.

[20] O. Lassila, and S. Ralph 1999. Resource Description Framework (RDF) Model and Syntax Specification. W3C available at http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[21] LOD project: Linking Open Data (LOD). W3C, available at http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData#Project_Description.

[22] Linking Open Drug Data (LODD) 2009. HCLSIG, W3C, available at http://www.w3.org/wiki/HCLSIG/LODD.

[23] D. A. Natale, C. N. Arighi, W. C. Barker, J. Blake, T. C. Chang, Z. Hu, H. Liu, B. Smith and C. H. Wu.Framework for a protein ontology. *BMC Bioinformatics,* 8 2009, Suppl 9**,** S1.

[24] A. Passant 2009. SIOC, SIOC Types and Health Care and Life Sciences. W3C, available at http://www.w3.org/TR/hcls-sioc/.

[25] A. Passant and P. Ciccarese 2009. SWAN/SIOC: Alignment Between the SWAN and SIOC Ontologies. W3C, available at http://www.w3.org/TR/hcls-swansioc/.

[26] E. Prud'hommeaux and A. Seaburn 2006. SPARQl Query Language for RDF. W3C, available at http://www.w3.org/TR/rdf-sparql-query/.

[27] E. Prud'hommeaux, H. Deus and M. S. Marshall. Tutorial: Query Federation with SWObjects. *In:* SWAT4LS, 2010. Available from *Nature Preceedings*

[28] N. Redaschi and UniProt Consortium. UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web'. *In:* 3rd International Biocuration Conference, 2009.

[29] A. Ruttenberg, J. A. Rees, M. Samwald, and M. S. Marshall. Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in Bioinformatics,* 10 (2009)**,** 193-204.

[30] M. Samwald, and H. Stenzhorn Simple, ontology-based representation of biomedical statements through fine-granular entity tagging and new web standards. *Journal of Biomedical Semantics* 2010, 1 Suppl 1:S5.

[31] L. Sauermann and R. Cyganiak 2008). Cool URIs for the Semantiuc Web. W3C, available at http://www.w3.org/TR/cooluris/

[32] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheurmann, N. Shah, P. L. Whetzel and S. LEWIS. The OBO Foundry: coordinated

evolution of ontologies to support biomedical data integration. *Nature Biotechnology,* 25 (2007)**,** 1251-5.

[33] Semantic Web Education and Outreach (SWEO) Interest Group. W3C, available at http://www.w3.org/blog/SWEO/.

[34] M. S. Marshall, B. Richard, H. Deus, J. Zhao, E. L. Willighagen, M. Samwald, E. Pichler, J. Hajagos, E. Prud'hommeaux and S. Stephens. Emerging practices for mapping and linking life sciences data using RDF — a case series. WebSemantics: Science, Services and Agents on the World Wide Web, April 2012

[35] V. Momtchev, D. Peychev, T. Primov, G. Georgiev. Expanding the Pathway and Interaction Knowledge in Linked Life Data. *In:* International Semantic Web Challenge, 2009.

[36] V. Mironov, N. Seethappan, W. Blonde, E. Antezana, B. Lindi, and M. Kuiper. Benchmarking triple stores with biological data. Semantic Web Applications and Tools for Life Sciences (SWAT4LS). Berlin, Germany 2010.