Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data

Alison Callahan[1,*], José Cruz-Toledo[1,*], Peter Ansell[2], and Michel Dumontier[1]

[1] Department of Biology, Carleton University, Ottawa, Canada
{acallaha,jctoledo}@connect.carleton.ca,
michel_dumontier@carleton.ca
[2] eResearch Lab, School of ITEE, University of Queensland, Brisbane, Australia
ansell.peter@gmail.com

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

Keywords: Semantic Web, RDF, Linked Data, Life Sciences, SPARQL.

1 Introduction

With the advent of the World Wide Web, journals have increasingly augmented their peer-reviewed journal publications with downloadable experimental data. While the increase in data availability should be cause for celebration, the potential for biomedical discovery across all of these data is hampered by access restrictions, incompatible formats, lack of semantic annotation and poor connectivity between datasets [1]. Although organizations such as the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) have made great strides to extract, capture and integrate data, the lack of formal, machine-understandable semantics results in ambiguity in the data and the relationships between them. With over 1500 biological databases, it becomes necessary to implement a more sophisticated scheme to unify the representation of diverse biomedical data so that it becomes easier to integrate and explore [2]. Importantly, there is a fundamental need to capture the provenance of these data in a manner that will support experimental design and reproducibility in scientific research. Providing data also presents real practical challenges, including ensuring persistence, availability, scalability, and providing the right tools to facilitate data exploration including query formulation.

[*] These authors contributed equally to this work.

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7882, pp. 200–212, 2013.

The Resource Description Framework (RDF) provides an excellent foundation to build a unified network of linked data on the emerging Semantic Web. While an increasing number of approaches are being proposed to describe and integrate specific biological data [3-5], it is the lack of coordinated identification, vocabulary overlap and alternative formalizations that challenges the promise of large-scale integration [6]. Formalization of data into ontologies using the Web Ontology Language (OWL) have yielded interesting results for integration, classification, consistency checking and more effective query answering with automated reasoning [7-11]. However, these efforts build the ontology in support of the task and there is little guarantee that the formalization will accommodate future data or support new applications. Alternatively, integration of data may be best facilitated by independent publication of datasets and their descriptions and subsequent coordination into integrative ontologies or community standards. This approach provides maximum flexibility for publishing original datasets with publisher provided descriptors in that they are not constrained by limited standards, but provides a clear avenue for future integration into a number of alternative standards.

Bio2RDF is a well-recognized open-source project that provides linked data for the life sciences using Semantic Web technologies. Bio2RDF scripts convert heterogeneously formatted data (e.g. flat-files, tab-delimited files, dataset specific formats, SQL, XML *etc*.) into a common format – RDF. Bio2RDF follows a set of basic conventions to generate and provide Linked Data which are guided by Tim Berners-Lee's design principles [1], the Banff Manifesto [2] and the collective experience of the Bio2RDF community. Entities, their attributes and relationships are named using a simple convention to produce Internationalized Resource Identifiers (IRIs) while statements are articulated using the lightweight semantics of RDF Schema (RDFS) and Dublin Core. Bio2RDF IRIs are resolved through the Bio2RDF Web Application, a servlet that answers Bio2RDF HTTP requests by formulating SPARQL queries against the appropriate SPARQL endpoints.

Although several efforts for provisioning linked life data exist such as Neurocommons [12], LinkedLifeData [13], W3C HCLS[3], Chem2Bio2RDF [14] and BioLOD, Bio2RDF stands out for several reasons: i) Bio2RDF is open source and freely available to use, modify or redistribute, ii) it acts on a set of basic guidelines to produce syntactically interoperable linked data across all datasets, iii) does not attempt to marshal data into a single global schema, iv) provides a federated network of SPARQL endpoints and v) provisions the community with an expandable global network of mirrors that host RDF datasets. Thus, Bio2RDF uniquely offers a community-focused resource for creating and enhancing the quality of biomedical data on the Semantic Web.

Here, we report on a second coordinated release of Bio2RDF Release 2 (R2), which yields substantial increases in syntactic and semantic interoperability across refactored Bio2RDF datasets. We address the problem of IRI inconsistency arising from independently generated scripts through an API over a dataset registry

[1] http://www.w3.org/DesignIssues/Principles.html

[2] https://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff_Manifesto

[3] http://www.w3.org/blog/hcls/

to generate validated IRIs. We further generate provenance and statistics for each dataset, and provide public SPARQL endpoints, downloadable database files and RDF files. We demonstrate federated SPARQL queries within and across the Bio2RDF network, including queries that make use of the Semanticscience Integrated Ontology (SIO) [4], which provides a simple model with a rich set of relations to coordinate ontologies, data and services.

2 Methods

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

2.1 Entity Naming

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

`http://bio2rdf.org/`*`namespace`*`:`*`identifier`*

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

`http://bio2rdf.org/hgnc:9599`

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

`http://bio2rdf.org/`*`namespace`*`_resource:`*`identifier`*

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

`http://bio2rdf.org/ctd_resource:C112297D029597`

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

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

`http://bio2rdf.org/`*`namespace`*`_vocabulary:`*`identifier`*

[4] `http://code.google.com/p/semanticscience/wiki/SIO`

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

```
http://bio2rdf.org/homologene_vocabulary:HomoloGene_Group
```

while the taxonomic relation is specified with:

```
http://bio2rdf.org/homologene_vocabulary:has_taxid
```

2.2 Open Source Scripts

In 2012, we consolidated the set Bio2RDF open source[5] scripts into a single GitHub repository (bio2rdf-scripts[6]). GitHub facilitates collaborative development through project forking, pull requests, code commenting, and merging. Thirty PHP scripts, one Java program and a Ruby gem are now available for any use (including commercial), modification and redistribution by anyone wishing to generate BioRDF data, or to improve the quality of RDF conversions currently used in Bio2RDF.

2.3 Programmatically Accessible Resource Registry

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

2.4 Provenance

Bio2RDF scripts now generate provenance using the Vocabulary of Interlinked Datasets (VoID), the Provenance vocabulary (PROV) and Dublin Core vocabulary. As illustrated in Fig. 1, each item in a dataset is linked using void:inDataset to a provenance object (typed as void:Dataset). The provenance object represents a Bio2RDF dataset, in that it is a version of the source data whose attributes include a label, the creation date, the creator (script URL), the publisher (Bio2RDF.org), the Bio2RDF license and rights, the download location for the dataset and the SPARQL endpoint in which the resource can be found. Importantly, we use the W3C PROV relation 'wasDerivedFrom' to link this Bio2RDF dataset to the source dataset, along with its licensing and source location.

---

[5] http://opensource.org/licenses/MIT

[6] https://github.com/bio2rdf/bio2rdf-scripts

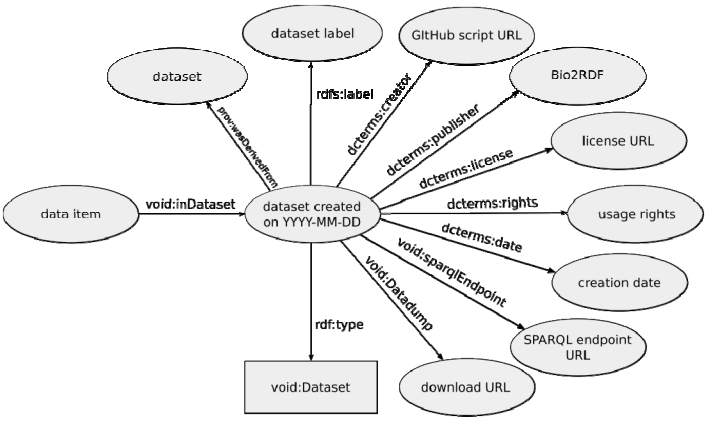

Fig. 1. The Bio2RDF R2 provenance model

2.5 Dataset Metrics

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

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

These metrics are serialized as RDF using our own vocabulary using the namespace `http://bio2rdf.org/dataset_vocabulary`), and subsequently loaded into a named graph at each dataset SPARQL endpoint with the following pattern:

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

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

```
PREFIX statistics: <http://bio2rdf.org/dataset_vocabulary:>
SELECT *
FROM <http://bio2rdf.org/bio2rdf-ctd-statistics>
WHERE {
 ?endpoint a statistics:Endpoint.
```

```
 ?endpoint statistics:has_type_relation_type_count ?c.
  ?c statistics:has_subject_type ?subjectType.
  ?c statistics:has_subject_count ?subjectCount.
  ?c statistics:has_predicate ?predicate.
  ?c statistics:has_object_type ?objectType.
  ?c statistics:has_object_count ?objectCount.
}
```

Furthermore, to support context-sensitive SPARQL query formulation using SparQLed [15], we generated the data graph summaries using the Dataset Analytics Vocabulary[7]. These are stored in each endpoint in the graph named http://sindice.com/analytics .

2.6 Bio2RDF to SIO Ontology Mapping

Since each Bio2RDF dataset is expressed in terms of a dataset-specific vocabulary for its types and relations, it becomes rather challenging to compose federated queries across both linked datasets as well as datasets that overlap in their content. To facilitate dataset-independent querying, Bio2RDF dataset-specific vocabulary were mapped to the Semanticscience Integrated Ontology (SIO), which is also being used to map vocabularies used to describe SADI-based semantic web services. Dataset specific types and relations were extracted using SPARQL queries and manually mapped to corresponding SIO classes, object properties and datatype properties using the appropriate subclass relation (i.e. rdfs:subClassOf, owl:SubObjectPropertyOf). Bio2RDF dataset vocabularies and their SIO-mappings are stored in separate OWL ontologies on the bio2rdf-mapping GitHub repository[8].

2.7 SPARQL Endpoints

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

2.8 Bio2RDF Web Application

Bio2RDF Linked Data IRIs are made resolvable through the Bio2RDF Web Application, a servlet based application that uses the QueryAll Linked Data library [16] to dynamically answer requests for Bio2RDF IRIs by aggregating the results of SPARQL queries to Bio2RDF SPARQL endpoints that are automatically selected based on the structure of the query IRI. The Web Application can be configured to resolve queries using multiple SPARQL endpoints, each of which may handle different namespaces and identifier patterns. Such configurations are stored as RDF, and specified using Web Application profiles. Profiles are designed to allow different

[7] http://vocab.sindice.net/analytics#
[8] https://github.com/bio2rdf/bio2rdf-mapping

hosts to reuse the same configuration documents in slightly different ways. For example, the Bio2RDF Web Application R2 profile has been configured to resolve queries that include the new '_resource' and '_vocabulary' namespaces (section 2.1), as well existing query types used by the base Bio2RDF profile, and to resolve these queries using the R2 SPARQL endpoints.

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

2.9 Resolving Bio2RDF IRIs Using Multiple SPARQL Endpoints

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

```
http://bio2rdf.org/uniprot:P05067
```

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

3 Results

3.1 Bio2RDF Release 2

Nineteen datasets, including 5 new datasets, were generated as part of R2 (**Table 1**). R2 also includes 3 datasets that are themselves aggregates of datasets which are now available as one resource. For instance, iRefIndex consists of 13 datasets (BIND, BioGRID, CORUM, DIP, HPRD, InnateDB, IntAct, MatrixDB, MINT, MPact, MPIDB, MPPI and OPHID) while NCBO's Bioportal collection currently consists of 100 OBO ontologies including ChEBI, Protein Ontology and the Gene Ontology.

We also have 10 additional updated scripts that are currently generating updated datasets and SPARQL endpoints to be available with the next release: ChemBL, DBPedia, GenBank, PathwayCommons, the RCSB Protein Databank, PubChem, PubMed, RefSeq, UniProt (including UniRef and UniParc) and UniSTS.

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

Table 1. Bio2RDF Release 2 datasets with select dataset metrics. The asterisks indicate datasets that are new to Bio2RDF.

Dataset	*Namespace*	*# of triples*	*# of unique subjects*	*# of unique predicates*	*# of unique objects*
Affymetrix	affymetrix	44469611	1370219	79	13097194
Biomodels*	biomodels	589753	87671	38	209005
Bioportal*	bioportal	15384622	4425342	191	7668644
Comparative Toxicogenomics Database	ctd	141845167	12840989	27	13347992
DrugBank	drugbank	1121468	172084	75	526976
NCBI Gene	ncbigene	394026267	12543449	60	121538103
Gene Ontology Annotations	goa	80028873	4710165	28	19924391
HUGO Gene Nomenclature Committee	hgnc	836060	37320	63	519628
Homologene	homologene	1281881	43605	17	1011783
InterPro*	interpro	999031	23794	34	211346
iProClass	iproclass	211365460	11680053	29	97484111
iRefIndex	irefindex	31042135	1933717	32	4276466
Medical Subject Headings	mesh	4172230	232573	60	1405919
National Drug Code Directory*	ndc	17814216	301654	30	650650
Online Mendelian Inheritance in Man	omim	1848729	205821	61	1305149
Pharmacogenomics Knowledge Base	pharmgkb	37949275	5157921	43	10852303
SABIO-RK*	sabiork	2618288	393157	41	797554
Saccharomyces Genome Database	sgd	5551009	725694	62	1175694
NCBI Taxonomy	taxon	17814216	965020	33	2467675
Total	**19**	**1,010,758,291**	**57850248**	**1003**	**298470583**

3.2 Metric Informed Querying

Dataset metrics (section 2.5) provide an overview of the contents of a dataset and can be used to guide the development of SPARQL queries. **Table 2** shows values for the type-relation-type metric in the DrugBank dataset. In the first row we note that 11,512 unique pharmaceuticals are paired with 56 different units using the 'form' predicate, indicating the enormous number of possible formulations. Further in the list, we see that 1,074 unique drugs are involved in 10,891 drug-drug interactions, most of these arising from FDA drug product labels.

Table 2. Selected DrugBank dataset metrics describing the frequencies of type-relation-type occurrences. The namespace for subject types, predicates, and object types is http://bio2rdf.org/drugbank_vocabulary

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

The type-relation-type metric gives the necessary information to understand how object types are related to one another in the RDF graph. It can also inform the construction of an immediately useful SPARQL query, without losing time generating 'exploratory' queries to become familiar with the dataset model. For instance, the above table suggests that in order to retrieve the targets that are involved in drug-target interactions, one should specify the 'target' predicate, to link to a target from its drug-target interaction(s):

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

Some of the results of this query are listed in Table 3.

Table 3. Partial results from a query to obtain drug-target interactions from the Bio2RDF DrugBank SPARQL endpoint

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

Dataset metrics can also facilitate federated queries over multiple Bio2RDF endpoints in a similar manner. For example, the following query retrieves all biochemical reactions from the Bio2RDF Biomodels endpoint that are kinds of "protein catabolic process", as defined by the Gene Ontology in the NCBO Bioportal endpoint:

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

3.3 Bio2RDF Dataset Vocabulary-SIO Mapping

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

```
define input:inference "http://bio2rdf.org/sio_mappings"
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX ctd_vocab: <http:bio2rdf.org/ctd_vocabulary:>
PREFIX ndc_vocab: <http://bio2rdf.org/ndc_vocabulary:>
SELECT ?chemical ?chemicalLabel
WHERE {
  #SIO_01126: 'chemical substance'
  ?chemical a sio:SIO_011126.
  ?chemical rdfs:label ?chemicalLabel .
  #affects Diabetes mellitus pathway
  ?chemical ctd_vocab:pathway <http://bio2rdf.org/kegg:04930> .
  #dosage form: tablet, extended release
  ?chemical ndc_vocab:dosage-form
  <http://bio2rdf.org/ndc_vocabulary:00426c812b33febc3f9cd1fee8
  cc83ce> .
}
```

This query is possible because the classes 'ctd_vocab:Chemical' and 'ndc_vocab:human-prescription-drug' have been mapped as subclasses of the SIO class 'chemical substance'[9].

4 Discussion

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

A major aspect of what makes Bio2RDF successful from a Linked Data perspective is the use of a central registry of datasets in order to normalize generated IRIs. Although we previously created a large aggregated namespace directory, the lack of extensive curation meant that the directory contained significant overlap and omissions. Importantly, no script specifically made use of this registry, and thus adherence to the namespaces was strictly in the hands of developers at the time of writing the code. In consolidating the scripts, we found significant divergence in the use of a preferred namespace for generating Bio2RDF IRIs, either because of the overlap in

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

directory content, or in the community adopting another preferred prefix. With the addition of an API to automatically generate the preferred Bio2RDF IRI from any number of dataset prefixes (community-preferred synonyms can be recorded), all Bio2RDF IRIs can be validated such that unknown dataset prefixes must be defined in the registry. Importantly, our registry has been shared with maintainers of identifiers.org in order for their contents to be incorporated into the MIRIAM registry [17] which powers that URL resolving service. Once we have merged our resource listings, we expect to make direct use of the MIRIAM registry to list new entries, and to have identifiers.org list Bio2RDF as a resolver for most of its entries. Moreover, since the MIRIAM registry describes regular expressions that specify the identifier pattern, Bio2RDF scripts will be able to check whether an identifier is valid for a given namespace, thereby improving the quality of data produced by Bio2RDF scripts.

The dataset metrics that we now compute for each Bio2RDF dataset have significant value for users and providers. First, users can get fast and easy access to basic dataset metrics (number of triples, *etc*.) as well as more sophisticated summaries such as which types are in the dataset and how are they connected to one another. This data graph summary is the basis for SparQLed, an open source tool to assist in query composition through context-sensitive autocomplete functionality. Use of these summaries also reduces the server load for data provider servers, which in turns frees up resources to more quickly respond to interesting domain-specific queries. Second, we anticipate that these metrics may be useful in monitoring dataset flux. Bio2RDF now plans to provide bi-annual release of data, and as such, we will develop infrastructure to monitor change in order to understand which datasets are evolving, and how are they changing. Thus, users will be better able to focus in on content changes and providers will be able to make informed decisions about the hardware and software resources required to provision the data to Bio2RDF users.

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

Acknowledgements. This research was supported by an NSERC CGSD to AC, and NSERC funding to JCT and MD. We also acknowledge technical support from Marc-Alexandre Nolin, constructive but anonymous peer-reviewers, and useful discussions from the Bio2RDF community.

References

1. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St. Pierre, S., et al.: Big data: The future of biocuration. Nature 455(7209), 47–50 (2008)
2. Goble, C., Stevens, R.: State of the nation in data integration for bioinformatics. J. Biomed. Inform. 41(5), 687–693 (2008)
3. Cerami, E.G., Bader, G.D., Gross, B.E., Sander, C.: cPath: open source software for collecting, storing, and querying biological pathways. BMC Bioinformatics 7, 497 (2006)

4. Chen, H., Yu, T., Chen, J.Y.: Semantic Web meets Integrative Biology: a survey. Brief Bioinform. (2012)
5. Ruebenacker, O., Moraru, I.I., Schaff, J.C., Blinov, M.L.: Integrating BioPAX pathway knowledge with SBML models. IET Syst. Biol. 3(5), 317–328 (2009)
6. Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., et al.: Toward interoperable bioscience data. Nat. Genet. 44(2), 121–126 (2012)
7. Berlanga, R., Jimenez-Ruiz, E., Nebot, V.: Exploring and linking biomedical resources through multidimensional semantic spaces. BMC Bioinformatics 13(suppl. 1), S6 (2012)
8. Gennari, J.H., Neal, M.L., Galdzicki, M., Cook, D.L.: Multiple ontologies in action: composite annotations for biosimulation models. J. Biomed. Inform. 44(1), 146–154 (2011)
9. Hoehndorf, R., Dumontier, M., Gennari, J.H., Wimalaratne, S., de Bono, B., Cook, D.L., Gkoutos, G.V.: Integrating systems biology models and biomedical ontologies. BMC Syst. Biol. 5, 124 (2011)
10. Hoehndorf, R., Dumontier, M., Oellrich, A., Rebholz-Schuhmann, D., Schofield, P.N., Gkoutos, G.V.: Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. PLoS One 6(7), e22006 (2011)
11. Jonquet, C., Lependu, P., Falconer, S., Coulet, A., Noy, N.F., Musen, M.A., Shah, N.H.: NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. Web Semant. 9(3), 316–324 (2011)
12. Ruttenberg, A., Rees, J.A., Samwald, M., Marshall, M.S.: Life sciences on the Semantic Web: the Neurocommons and beyond. Brief Bioinform. 10(2), 193–204 (2009)
13. Momtchev, V., Peychev, D., Primov, T., Georgiev, G.: Expanding the Pathway and Interaction Knowledge in Linked Life Data. In: Semantic Web Challenge: 2009, Amsterdam (2009)
14. Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 11, 255 (2010)
15. Campinas, S., Perry, T.E., Ceccarelli, D., Delbru, R., Tummarello, G.: Introducing RDF Graph Summary with Application to Assisted SPARQL Formulation, pp. 261–266 (2012)
16. Ansell, P.: Model and prototype for querying multiple linked scientific datasets. Future Generation Computer Systems 27(3), 329–333 (2011)
17. Juty, N., Le Novere, N., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 40(Database issue), D580–D586 (2012)