Improved dataset coverage and interoperability with Bio2RDF Release 2

Alison Callahan[1], Jose Cruz-Toledo[1], Peter Ansell[2], Dana Klassen[3], Giovanni Tummarello[4] and Michel Dumontier[1§]

[1]Department of Biology, Carleton University, Ottawa, Canada, [2]Microsoft QUT eResearch Centre, Queensland University of Technology, Australia, [3]Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland [4]SindiceTech, Galway, Ireland
[§]Corresponding author

Abstract. Bio2RDF is an open source project that uses Semantic Web technologies to create and provide the largest network of Linked Data for the life sciences. Here, we present the second release of the Bio2RDF project which features updated, open-source scripts, a resource registry for IRI mapping and normalization, dataset provenance, data metrics, downloadable RDF data files and Virtuoso SPARQL endpoints. We describe dataset connectivity, assisted SPARQL queries with context-aware SPARQLed, and mashup capability using the Sig.ma search engine. We discuss updates to the Bio2RDF project in the context of other related resources as well as future improvements.

Keywords: semantic web, linked data, life sciences data, SPARQL

1 Introduction

In this post-genomic-information era, biological researchers are often confronted with the inevitable and unenviable task of having to integrate their experimental results with those of others. This task usually involves a tedious manual search and assimilation of often isolated and diverse collections of life sciences data hosted by multiple independent providers including organizations such as the National Center for Biotechnology Information (NCBI)[1] and the European Bioinformatics Institute (EBI)[2] which provide dozens of user-submitted and curated data, as well as smaller institutions such as the Donaldson group which publishes iRefIndex[1], a database of molecular interactions aggregated from 13 data sources. While these mostly isolated silos of biological information occasionally provide links between their records (*e.g.* UniProt links its entries to hundreds of other databases[3]), they are typically serialized in either HTML tags or in flat file data dumps that lack the semantic richness required to serialize the intent of the linkage between data records. With thousands of biological

[1] http://www.ncbi.nlm.nih.gov/
[2] http://www.ebi.ac.uk/
[3] http://www.uniprot.org/database/

databases[4,5] and hundreds of thousands if not millions of datasets, our ability to find relevant data is hampered by non-standard database interfaces and an enormous number of haphazard data formats[2]. Moreover, metadata about these biological data providers (dataset source data information, dataset versioning, licensing information, date of creation, *etc.*) is often difficult to obtain. Taken together, our inability to easily navigate through available data presents an overwhelming barrier to their reuse.

Bio2RDF is an open source project that uses Semantic Web technologies to make possible the distributed querying of integrated life sciences data. Since its inception[3], Bio2RDF has made use of the Resource Description Framework (RDF) and the RDF Schema (RDFS) to unify the representation of data obtained from diverse (molecules, enzymes, pathways, diseases, *etc.*) and heterogeneously formatted biological data (*e.g.* flat-files, tab-delimited files, SQL, dataset specific formats, XML *etc.*). Once converted to RDF, this biological data can be queried using the powerful SPARQL Protocol and RDF Query Language (SPARQL), which can be used to federate queries across multiple SPARQL-compliant databases (a.k.a. SPARQL endpoints).

Although several efforts for provisioning linked life data exist such as Neurocommons[4], LinkedLifeData[5], W3C HCLS[6], Chem2Bio2RDF[6] and BioLOD[7], 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 Bio2RDF datasets.

Here we present Bio2RDF Release 2, a significant update from past practice that considerably increases the level of syntactic interoperability across datasets through a script-directed IRI normalization that queries a central dataset registry. We also introduce a new model for data item-level provenance and describe new metrics for linked datasets that guide querying and provide high-level descriptions of datasets. We characterize dataset connectivity, assisted SPARQL queries with context-aware SPARQLed, and data mash-up capability using the Sig.ma[8] search engine.

2 Methods

2.1 Resource Registry

A resource registry composed of vocabularies (*e.g.* Gene Ontology, ChEBI, *etc.*) and datasets (*e.g.* RefSeq) was developed to facilitate dataset identification and inter-dataset mapping. Each item lists a preferred short name (a.k.a. namespace; *e.g.* ‘pdb’

[4] http://nar.oxfordjournals.org/content/40/D1.toc

[5] http://www.freebase.com/view/base/bio2rdf/views/bm

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

[7] http://biolod.org/

[8] http://sig.ma/

for the Protein DataBank), resource synonyms (*e.g.* ncbigene, entrez gene, entrez-gene/locuslink for the NCBI's Gene database), as well as primary and secondary base Internationalized Resource Identifiers (IRIs) used within the datasets (*e.g.*http://purl.obolibrary.org/obo/, http://purl.org/obo/owl/, http://purl.obofoundry.org/namespace, *etc*). The resource registry is currently available as part of the PHP-LIB project[9].

2.2 Identifiers

Bio2RDF data items are identified by formulating an Internationalized Resource Identifier (IRI) consisting of the following pattern:

http://bio2rdf.org/namespace:identifier

where 'namespace' is the preferred short name of a biological dataset as found in the resource registry (section 2.1) and the 'identifier' is the unique string used by the source provider. For example, the Protein DataBank (PDB) features a structure containing an adenine riboswitch complex, which it identifies by the accession "1Y26". In the registry, the PDB is assigned the namespace "pdb" and thus, its corresponding Bio2RDF IRI is

http://bio2rdf.org/pdb:1Y26

Two additional identifier patterns are used for resources introduced as a product of RDFization. First, *namespace_vocabulary:identifier*, is used to name dataset-specific types and predicates. For example, the chemoinformatics resource DrugBank contains data about drugs and their targets, and these two types have the following IRIs:

http://bio2rdf.org/drugbank_vocabulary:Drug
http://bio2rdf.org/drugbank_vocabulary:Target

The second namespace pattern, *namespace_resource:identifier,* is used to designate additional resources that were introduced to convert (unidentified) n-ary relations into an identified object with a set of binary relations. For example, the Pharmacogenomics Knowledge Base (PharmGKB) describes associations between diseases, genes and drugs, but does not specify an identifier for either of these associations, and hence we assign a new stable identifier for each, such as

http://bio2rdf.org/pharmgkb_resource:association_PA445019_PA126

for the gene-disease association between cytochrome P450, family 2, subfamily C, polypeptide 9 (pharmgkb:PA126) and Myocardial Infarction (pharmgkb:PA445019).

[9] https://github.com/micheldumontier/php-lib/blob/master/ns.php

2.3 Bio2RDF's Open Scripts

At its core, Bio2RDF is a set of conventions to generate and provide Linked Data. These best practices have been inspired by the Banff Manifesto[10], Tim Berner-Lee's design principles[11] and the collective experience of the Bio2RDF community. In 2012, we consolidated the set Bio2RDF open source[12] scripts into a single GitHub repository (bio2rdf-scripts)[13], which 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 RDF data on their own, or to improve the quality of RDF conversions currently used in Bio2RDF.

Nearly every script has now been updated to make use of the resource registry, thereby ensuring a high level of syntactic interoperability between the generated linked data sets. Scripts that have not yet been updated include the NCBO Bioportal collection, GenBank and RefSeq. These transformation scripts are programmatically restricted to only create valid Bio2RDF resources and only make use of preferred namespace items in a dataset as found in our resource registry.

2.4 Provenance

Previous iterations of Bio2RDF scripts lacked a framework with which to record provenance (metadata about the creator, creation date and origin) for Bio2RDF datasets. Upon execution, Bio2RDF scripts now generate provenance records using the W3C Vocabulary of Interlinked Datasets (VoID), the Provenance vocabulary (PROV) and Dublin Core vocabulary. Each data item is linked to a provenance object that indicates the source of the data, the time at which the RDF was generated, licensing (if available from data source provider), the SPARQL endpoint in which the resource can be found, and the downloadable RDF file where the data item is located. Each dataset provenance object has a unique IRI and label based on the dataset name and creation date. The date-specific dataset IRI is linked to a unique dataset IRI using the W3C PROV predicate 'wasDerivedFrom' such that one can query the dataset SPARQL endpoint to retrieve all provenance records for datasets created on different dates. **Figure 1** shows an example provenance record for the NLM Medical Subject Headings (MeSH) dataset. Each resource in the dataset is linked the date-unique dataset IRI that is part of the provenance record using the VoID 'inDataset' predicate. Other important features of the provenance record include the use of the Dublin Core 'creator' term to link a dataset to the script on Github that was used to generate it, the VoID predicate 'sparqlEndpoint' to point to the dataset SPARQL endpoint, and VoID predicate 'dataDump' to point to the data download URL.

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

[11] http://www.w3.org/DesignIssues/LinkedData.html

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

[13] http://github.com/bio2rdf/bio2rdf-scripts

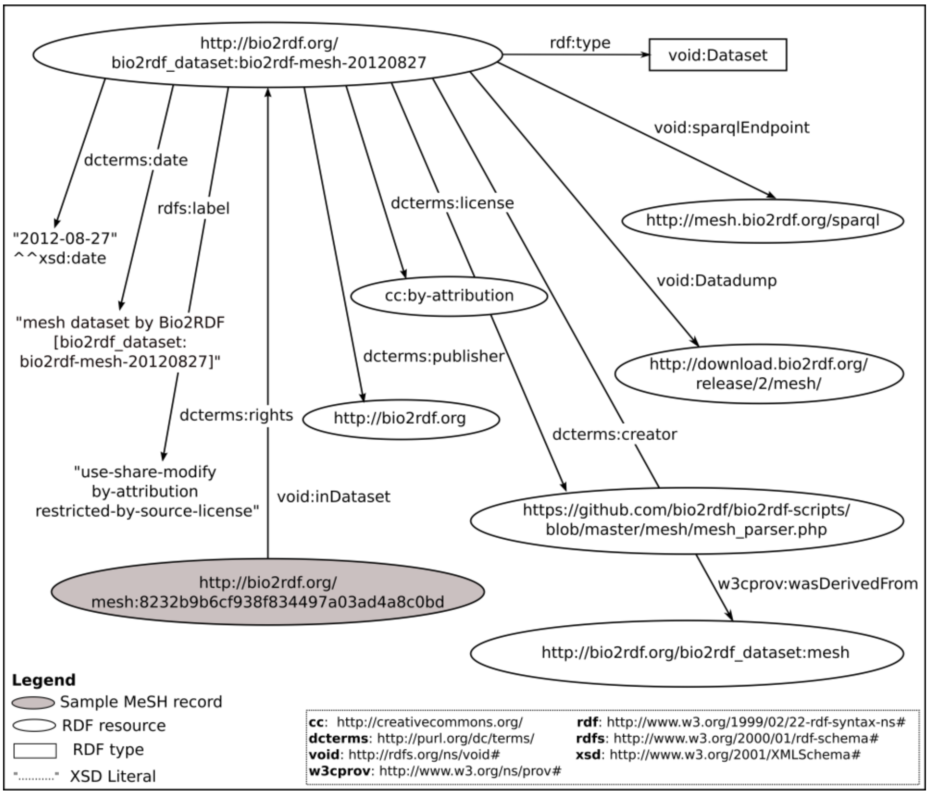

http://bio2rdf.org/
bio2rdf_dataset:bio2rdf-mesh-20120827
rdf:type
void:Dataset
void:sparqlEndpoint
http://mesh.bio2rdf.org/sparql
dcterms:date
"2012-08-27"
^^xsd:date
rdfs:label
"mesh dataset by Bio2RDF
[bio2rdf_dataset:
bio2rdf-mesh-20120827]"
dcterms:license
cc:by-attribution
void:Datadump
http://download.bio2rdf.org/
release/2/mesh/
dcterms:publisher
http://bio2rdf.org
dcterms:rights
"use-share-modify
by-attribution
restricted-by-source-license"
void:inDataset
dcterms:creator
https://github.com/bio2rdf/bio2rdf-scripts/
blob/master/mesh/mesh_parser.php
http://bio2rdf.org/
mesh:8232b9b6cf938f834497a03ad4a8c0bd
w3cprov:wasDerivedFrom
http://bio2rdf.org/bio2rdf_dataset:mesh
Legend
Sample MeSH record
RDF resource
RDF type
"..........." XSD Literal
cc: http://creativecommons.org/
dcterms: http://purl.org/dc/terms/
void: http://rdfs.org/ns/void#
w3cprov: http://www.w3.org/ns/prov#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
xsd: http://www.w3.org/2001/XMLSchema#

Figure 1 Example provenance record for the MeSH dataset

2.5 SPARQL Endpoints

Each dataset was loaded into a separate instance of OpenLink Virtuoso Community Edition build 06.01.3127 with the faceted browser, SPARQL 1.1 query federation and Cross-Origin Resource Sharing enabled.

2.6 Dataset metrics

Dataset metrics provide an important overview of dataset contents, which can be used to support query formulation or monitor changes to datasets over time. We apply three different dataset metrics programs (A-C below) to each dataset. These metrics are serialized as RDF and loaded into their own graphs at each dataset SPARQL endpoint.

A) Nine dataset metrics are computed[14] using SPARQL queries that obtain the following information

1. total number of triples
2. number of unique subjects
3. number of unique predicates
4. number of unique objects
5. number of unique types
6. unique predicate-object links and their frequencies
7. unique predicate-literal links and their frequencies
8. unique subject type-predicate-object type links and their frequencies
9. unique subject type-predicate-literal links and their frequencies

B) Namespace-related metrics are tabulated including

1. total number of references to a namespace
2. total number of inter-namespace references
3. total number of inter-namespace-predicate references

C) Data graph summaries[7] required for query formulation using SparQLed[15] are generated. The data graph summaries include metrics regarding the frequency and relationship among types via predicates. The data graph summaries are serialized in RDF using the Dataset Analytics Vocabulary[16].

3 Results

3.1 Bio2RDF Release 2

Nineteen datasets, including 5 new datasets, were generated as part of the Bio2RDF 2 release (**Table 1**). Several of the new datasets are themselves collections of datasets that 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: UniProt (including UniRef and UniParc), UniSTS, PubMed, PDB, RefSeq, PubChem, ChemBL, DBPedia, GenBank, MGI and PathwayCommons. Several of these datasets are the most resource intensive to generate and load, hence their later release schedule.

Each dataset has been loaded into a dataset specific SPARQL endpoint using Openlink Virtuoso version 6.1.6. SPARQL endpoints are available at

[14] https://github.com/bio2rdf/bio2rdf-scripts/blob/master/statistics/bio2rdf_stats_virtuoso.php
[15] https://github.com/sindicetech/sparqled
[16] http://vocab.sindice.net/analytics#

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. Pre-Release 2 Bio2RDF datasets are also available for download.

Table 1. Bio2RDF Release 2 datasets and selected dataset metrics. Dataset names annotated with * are new to the Bio2RDF network.

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

3.2 Namespace-based dataset connectivity

Figure 2 shows the connectivity between Bio2RDF datasets based on namespace-namespace linkages. Highlighted are core Bio2RDF datasets that make reference to hundreds of other datasets.

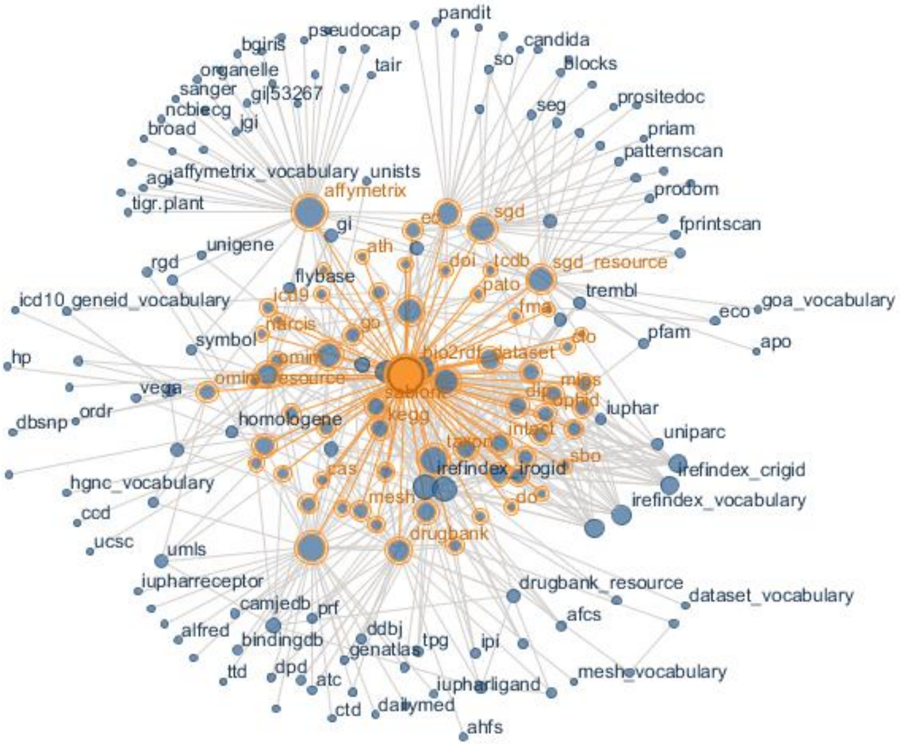


Figure 2 A network-based visualization of Bio2RDF namespace connectivity. Selected nodes indicate Bio2RDF datasets, as identified from provenance descriptions. Figure produced using IBM's Many Eyes (http://www-958.ibm.com).

3.3 Metrics-informed querying

Dataset metrics (section 2.6) serve as an overview of the contents of a dataset and can be used to guide querying with SPARQL. **Table 2** shows values for the type-relation-type metric in the DrugBank dataset. In the first row we observe 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 1074 unique drugs are involved in 10891 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 objects 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 drugs that are involved in drug-drug interactions, one should specify the 'ddi-interactor-in' predicate, to link a drug to its drug-drug interaction(s):

```
PREFIX drugbank_vocabulary: <http://bio2rdf.org/drugbank_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?ddi ?d1name
WHERE {
        ?ddi a drugbank_vocabulary:Drug-Drug-Interaction .
        ?d1 drugbank_vocabulary:ddi-interactor-in ?ddi .
        ?d1 rdfs:label ?d1name?.
        ?d2 drugbank_vocabulary:ddi-interactor-in ?ddi .
        ?d2 rdfs:label ?d2name.
       FILTER (?d1 != ?d2)
}
```

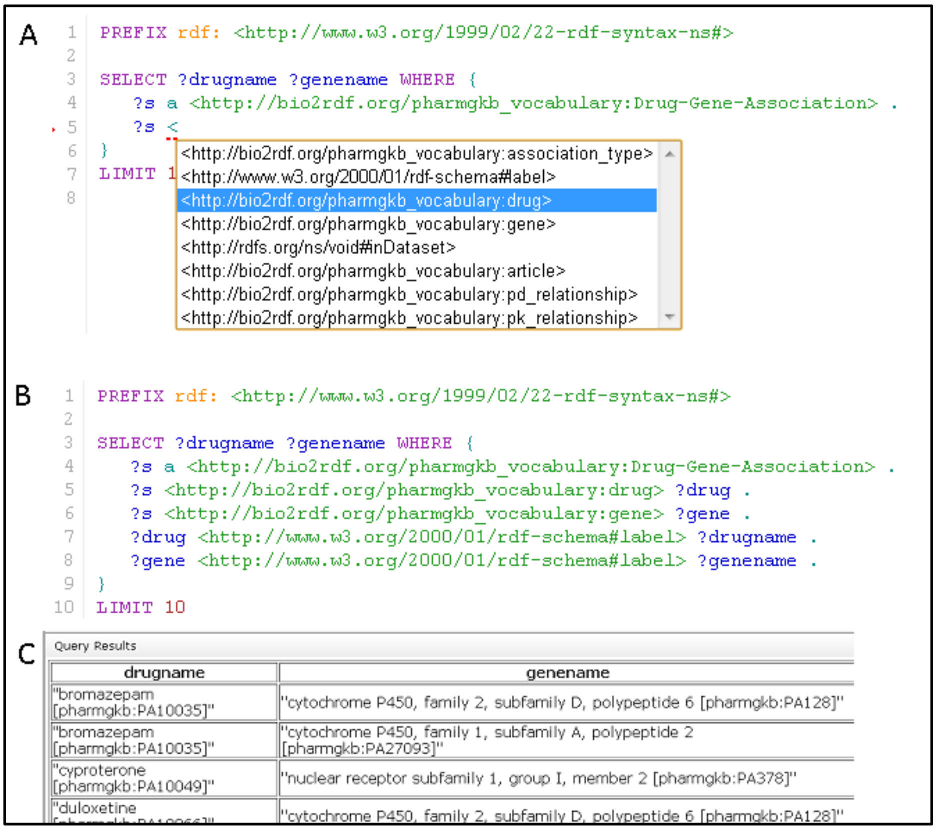
Some of the results of this query are listed in **Table 3**.

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

Drug-Drug Interaction	**DDI Drug Participants**
drugbank_resource:DB00001_DB01381	*Ginkgo biloba*, Lepirudin
drugbank_resource:DB00008_DB01223	Peginterferon alfa-2a, Aminophylline
drugbank_resource:DB00013_DB01404	Ginseng, Urokinase
drugbank_resource:DB00015_DB00208	Reteplase, Ticlopidine
drugbank_resource:DB00021_DB01409	Tiotropium, Secretin
drugbank_resource:DB00031_DB00055	Drotrecoginalfa, Tenecteplase
drugbank_resource:DB00041_DB01013	Aldesleukin, Clobetasol
drugbank_resource:DB00047_DB00195	Betaxolol, Insulin Glargine
drugbank_resource:DB00054_DB00775	Tirofiban, Abciximab
drugbank_resource:DB00059_DB00072	Trastuzumab, Betamethasone

3.4 Context-Aware SPARQL assistance with SPARQLed

SPARQLed is an open-source web-application that provides context sensitive IRI suggestions while formulating SPARQL queries. In particular, once a variable has been linked to a predicate or type, it is possible to deduce which other relations or types are applicable based on the inferred position of the object in the type-relation-type graph. **Figure 3** shows the grammar-sensitive and context aware formulation of a query to retrieve drug-gene associations from PharmGKB, where once the variable ?s is restricted to Drug-Gene-Association (Figure 3A) the only predicates available to use are listed in the suggestion box. Completion of the query (Figure 3B) to obtain the drug and gene names yields the results in Figure 3C.



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?drugname ?genename WHERE {
   ?s a <http://bio2rdf.org/pharmgkb_vocabulary:Drug-Gene-Association> .
   ?s <
}
LIMIT 1
```

<http://bio2rdf.org/pharmgkb_vocabulary:association_type>
<http://www.w3.org/2000/01/rdf-schema#label>
<http://bio2rdf.org/pharmgkb_vocabulary:drug>
<http://bio2rdf.org/pharmgkb_vocabulary:gene>
<http://rdfs.org/ns/void#inDataset>
<http://bio2rdf.org/pharmgkb_vocabulary:article>
<http://bio2rdf.org/pharmgkb_vocabulary:pd_relationship>
<http://bio2rdf.org/pharmgkb_vocabulary:pk_relationship>

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?drugname ?genename WHERE {
   ?s a <http://bio2rdf.org/pharmgkb_vocabulary:Drug-Gene-Association> .
   ?s <http://bio2rdf.org/pharmgkb_vocabulary:drug> ?drug .
   ?s <http://bio2rdf.org/pharmgkb_vocabulary:gene> ?gene .
   ?drug <http://www.w3.org/2000/01/rdf-schema#label> ?drugname .
   ?gene <http://www.w3.org/2000/01/rdf-schema#label> ?genename .
}
LIMIT 10
```

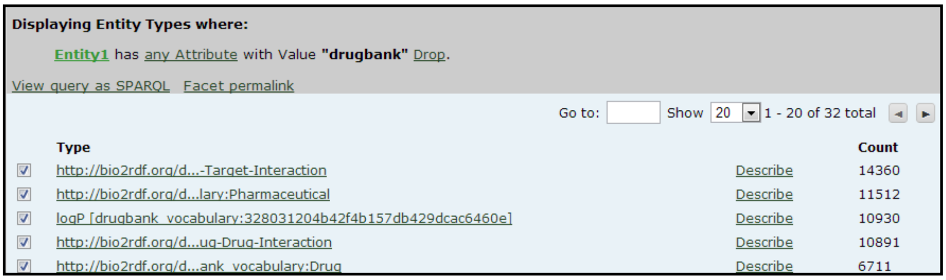
Query Results

drugname	genename
"bromazepam [pharmgkb:PA10035]"	"cytochrome P450, family 2, subfamily D, polypeptide 6 [pharmgkb:PA128]"
"bromazepam [pharmgkb:PA10035]"	"cytochrome P450, family 1, subfamily A, polypeptide 2 [pharmgkb:PA27093]"
"cyproterone [pharmgkb:PA10049]"	"nuclear receptor subfamily 1, group I, member 2 [pharmgkb:PA378]"
"duloxetine	"cytochrome P450, family 2, subfamily D, polypeptide 6 [pharmgkb:PA128]"

Figure 3 Using SPARQLed context-aware SPARQL assisted querying. (A) Selecting ctrl-shift space shows available predicates for a subject that has been constrained to a PharmGKB drug-gene association. (B) A SPARQLed-assisted query to get the drug and gene name. (C) First four drug-gene associations from the query in (B).

3.5 Virtuoso faceted search and query builder

By default, Virtuoso comes with a faceted browser that facilitates search and querying across a single SPARQL endpoint. The faceted search is initialized with a keyword (*e.g.* "drugbank" against the DrugBank endpoint – which appears in the rdfs:label of every drugbank resource). The search identifies 170,336 page-ranked hits that can be further categorized by type by selecting "Types" in the Entity Relations Navigation panel. The results include 32 types including drugs, drug interactions, targets, experimental and computed properties (**Figure 4**). Selecting any one of these will provide a list of specific instances of those types.



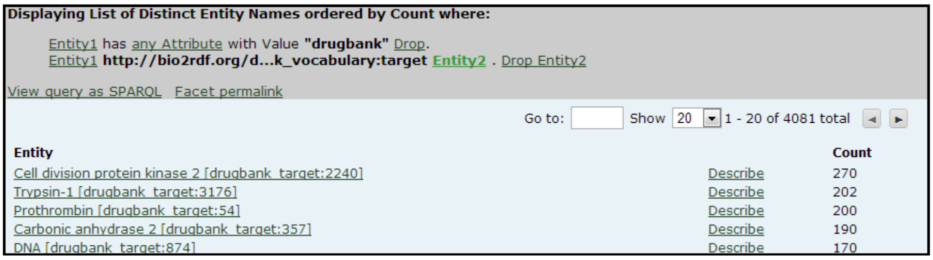
Displaying Entity Types where:

Entity1 has any Attribute with Value **"drugbank"** Drop.

	Type		Count
☑	http://bio2rdf.org/d...-Target-Interaction	Describe	14360
☑	http://bio2rdf.org/d...lary:Pharmaceutical	Describe	11512
☑	logP [drugbank_vocabulary:328031204b42f4b157db429dcac6460e]	Describe	10930
☑	http://bio2rdf.org/d...ug-Drug-Interaction	Describe	10891
☑	http://bio2rdf.org/d...ank_vocabulary:Drug	Describe	6711

Figure 4 Types matching a search of "drugbank" on the DrugBankVirtuoso endpoint.

However, the Virtuoso Faceted Search is significantly more powerful than just a search and navigation tool- it facilitates the iterative construction of an increasingly sophisticated query. For example, to determine the most popular target in DrugBank, first select the“attributes” link, which provides a list of predicates, including the drugbank_vocabulary:target, which points to DrugBank Targets. Selecting this predicate displays a list which can then be aggregated using “Distinct values (Aggregated)” to rank the targets by the number of entities that link to it using the ‘drugbank_vocabulary:target’ predicate. **Figure 5** shows that cell division protein kinase 2 is the highest referenced target (270 times) in DrugBank. Selecting “Entity1” in the top part of the query builder then shows the 270 drugs that target this enzyme, as well as the option to view the SPARQL query behind the faceted search and get a permalink to the facet.



Displaying List of Distinct Entity Names ordered by Count where:

Entity1 has any Attribute with Value **"drugbank"** Drop.
Entity1 **http://bio2rdf.org/d...k_vocabulary:target** Entity2 . Drop Entity2

View query as SPARQL   Facet permalink

Go to: □ Show 20 ▾ 1 - 20 of 4081 total ◄ ►

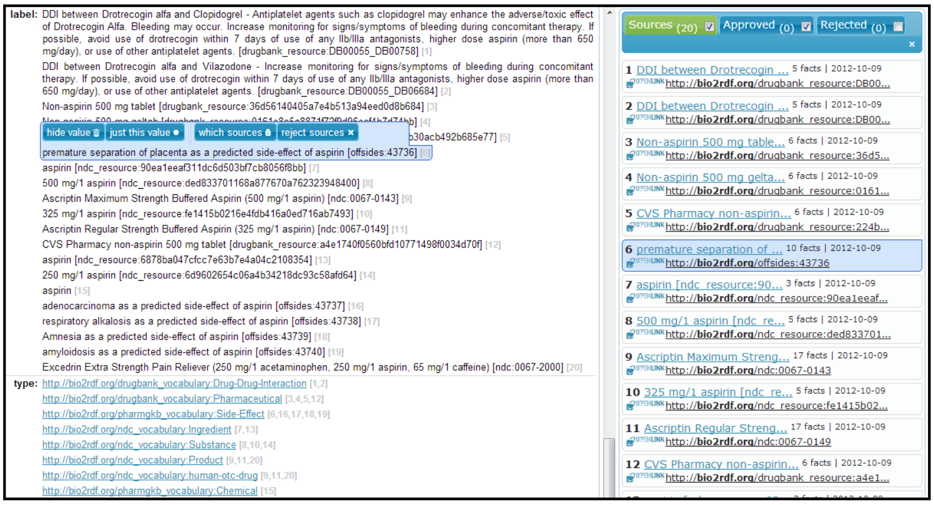
Entity		Count
Cell division protein kinase 2 [drugbank_target:2240]	Describe	270
Trypsin-1 [drugbank_target:3176]	Describe	202
Prothrombin [drugbank_target:54]	Describe	200
Carbonic anhydrase 2 [drugbank_target:357]	Describe	190
DNA [drugbank_target:874]	Describe	170

Figure 5 A count-ranked list of the attributes for all drug-target interactions.

3.6 Sig.ma powered mashups

Sig.ma[17] is an online browser that enables the mashup of data from one or more online resources (REST APIs, SPARQL endpoints, etc) using a keyword based search. We set up an instance of sig.ma to point to three endpoints (PharmGKB, DrugBank, NDC) and searched for ‘aspirin’. What is returned (**Figure 6**) is a mash-up of all resources that have “aspirin” in the rdfs:label, which is evident from the set of 23 labels and 8 types from the 3 endpoints (DrugBank: drug-drug interactions, pharmaceutical, side-effect; NDC: ingredient, substance, product and human OTC; PharmGKB: chemical).While having all the labels listed together is an unusual UI design, each attribute is linked to its source data item. By “approving” a source item, and hiding all the others, it becomes possible to see a single entry (**Figure 7**).

[17] http://sig.ma/



label: DDI between Drotrecogin alfa and Clopidogrel - Antiplatelet agents such as clopidogrel may enhance the adverse/toxic effect of Drotrecogin Alfa. Bleeding may occur. Increase monitoring for signs/symptoms of bleeding during concomitant therapy. If possible, avoid use of drotrecogin within 7 days of use of any IIb/IIIa antagonists, higher dose aspirin (more than 650 mg/day), or use of other antiplatelet agents. [drugbank_resource:DB00055_DB00758] [1]

DDI between Drotrecogin alfa and Vilazodone - Increase monitoring for signs/symptoms of bleeding during concomitant therapy. If possible, avoid use of drotrecogin within 7 days of use of any IIb/IIIa antagonists, higher dose aspirin (more than 650 mg/day), or use of other antiplatelet agents. [drugbank_resource:DB00055_DB06684] [2]

Non-aspirin 500 mg tablet [drugbank_resource:36d56140405a7e4b513a94eed0d8b684] [3]

Non-aspirin 500 mg geltab [drugbank_resource:0161e8e5a8871f72f9d96eef1b7d74bb] [4]

b30acb492b685e77] [5]

premature separation of placenta as a predicted side-effect of aspirin [offsides:43736] [6]

aspirin [ndc_resource:90ea1eeaf311dc6d503bf7cb8056f8bb] [7]

500 mg/1 aspirin [ndc_resource:ded833701168a877670a762323948400] [8]

Ascriptin Maximum Strength Buffered Aspirin (500 mg/1 aspirin) [ndc:0067-0143] [9]

325 mg/1 aspirin [ndc_resource:fe1415b0216e4fdb416a0ed716ab7493] [10]

Ascriptin Regular Strength Buffered Aspirin (325 mg/1 aspirin) [ndc:0067-0149] [11]

CVS Pharmacy non-aspirin 500 mg tablet [drugbank_resource:a4e1740f0560bfd10771498f0034d70f] [12]

aspirin [ndc_resource:6878ba047cfcc7e63b7e4a04c2108354] [13]

250 mg/1 aspirin [ndc_resource:6d9602654c06a4b34218dc93c58afd64] [14]

aspirin [15]

adenocarcinoma as a predicted side-effect of aspirin [offsides:43737] [16]

respiratory alkalosis as a predicted side-effect of aspirin [offsides:43738] [17]

Amnesia as a predicted side-effect of aspirin [offsides:43739] [18]

amyloidosis as a predicted side-effect of aspirin [offsides:43740] [19]

Excedrin Extra Strength Pain Reliever (250 mg/1 acetaminophen, 250 mg/1 aspirin, 65 mg/1 caffeine) [ndc:0067-2000] [20]

type: http://bio2rdf.org/drugbank_vocabulary:Drug-Drug-Interaction [1,2]
http://bio2rdf.org/drugbank_vocabulary:Pharmaceutical [3,4,5,12]
http://bio2rdf.org/pharmgkb_vocabulary:Side-Effect [6,16,17,18,19]
http://bio2rdf.org/ndc_vocabulary:Ingredient [7,13]
http://bio2rdf.org/ndc_vocabulary:Substance [8,10,14]
http://bio2rdf.org/ndc_vocabulary:Product [9,11,20]
http://bio2rdf.org/ndc_vocabulary:human-otc-drug [9,11,20]
http://bio2rdf.org/pharmgkb_vocabulary:Chemical [15]

1 DDI between Drotrecogin ... 5 facts | 2012-10-09
http://bio2rdf.org/drugbank_resource:DB00...

2 DDI between Drotrecogin ... 5 facts | 2012-10-09
http://bio2rdf.org/drugbank_resource:DB00...

3 Non-aspirin 500 mg table... 6 facts | 2012-10-09
http://bio2rdf.org/drugbank_resource:36d5...

4 Non-aspirin 500 mg gelta... 6 facts | 2012-10-09
http://bio2rdf.org/drugbank_resource:0161...

5 CVS Pharmacy non-aspirin... 6 facts | 2012-10-09
http://bio2rdf.org/drugbank_resource:224b...

6 premature separation of ... 10 facts | 2012-10-09
http://bio2rdf.org/offsides:43736

7 aspirin [ndc_resource:90... 3 facts | 2012-10-09
http://bio2rdf.org/ndc_resource:90ea1eeaf...

8 500 mg/1 aspirin [ndc_re... 5 facts | 2012-10-09
http://bio2rdf.org/ndc_resource:ded833701...

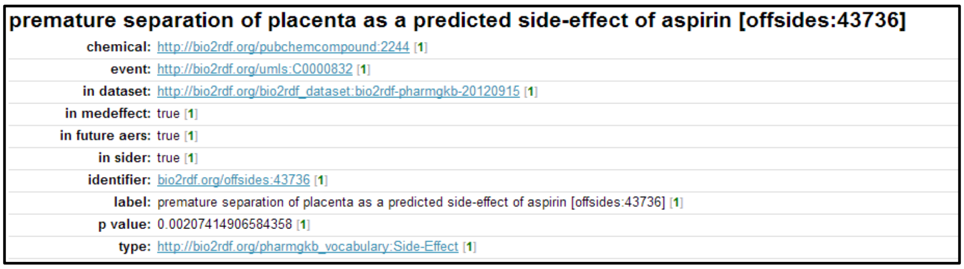
9 Ascriptin Maximum Streng... 17 facts | 2012-10-09
http://bio2rdf.org/ndc:0067-0143

10 325 mg/1 aspirin [ndc_re... 5 facts | 2012-10-09
http://bio2rdf.org/ndc_resource:fe1415b02...

11 Ascriptin Regular Streng... 17 facts | 2012-10-09
http://bio2rdf.org/ndc:0067-0149

12 CVS Pharmacy non-aspirin... 6 facts | 2012-10-09
http://bio2rdf.org/drugbank_resource:a4e1...

Figure 6 Sig.ma search with "aspirin" over PharmGKB, DrugBank and NDC



premature separation of placenta as a predicted side-effect of aspirin [offsides:43736]

chemical:	http://bio2rdf.org/pubchemcompound:2244 [1]
event:	http://bio2rdf.org/umls:C0000832 [1]
in dataset:	http://bio2rdf.org/bio2rdf_dataset:bio2rdf-pharmgkb-20120915 [1]
in medeffect:	true [1]
in future aers:	true [1]
in sider:	true [1]
identifier:	bio2rdf.org/offsides:43736 [1]
label:	premature separation of placenta as a predicted side-effect of aspirin [offsides:43736] [1]
p value:	0.00207414906584358 [1]
type:	http://bio2rdf.org/pharmgkb_vocabulary:Side-Effect [1]

Figure 7 View of a single entry from the sig.ma mashup

4 Discussion and Conclusions

Bio2RDF Release 2 features updates to data conversion scripts, datasets and functionality. The use of GitHub as an open software development environment makes it possible for enthusiasts to contribute new code and make improvements and suggestions to existing code. We welcome those that think Bio2RDF could be useful to their projects to contact us on the mailing list and participate in the development team.

The use of a Bio2RDF resource registry in each script will ensure that all Bio2RDF IRIs are in fact using validated namespaces (resource short names). Importantly, the addition of synonyms means that scripts can now map infrequently used or unusual database names and IRIs to a canonical Bio2RDF IRI. Our effort to develop a consistent registry of datasets and namespaces follows in the footsteps of our large scale aggregated namespace directory. Importantly, we have provided this directory to the maintainers of identifiers.org to be incorporated into the MIRIAM registry [8] which powers it. 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.

While we have described how dataset metrics are useful to summarize the RDF graph, and can be used to facilitate the construction of SPARQL queries as exemplified by the SPARQLed tool, we anticipate that these metrics will also be fundamentally useful in monitoring dataset flux. Users will no longer need to perform expensive queries over Bio2RDF endpoints to assess changes or updates to data as the relevant information (such as total number of triples, number of records of a given type, type-type relations *etc.*) is available in the pre-computed metrics, which will be generated with each data release and recorded as a 'snapshot' of the dataset at creation time. This is particularly timely, as recent efforts at the 2012 BioHackathon in Japan yielded an effort to assess the "sparkliness" of SPARQL endpoints[18] and to monitor their uptime. The dataset metrics also make it possible to assess the growth of datasets over time, in order to make projections about the hardware and software resources required to provision the data to Bio2RDF users. This will become increasingly important as we explore the provision of Bio2RDF data and related services in a cloud computing environment.

In summary, Bio2RDF Release 2 features updates to dataset conversion scripts as well as new datasets, a framework for recording dataset provenance, and a set of scripts to generate and publish Bio2RDF dataset metrics. We have demonstrated how multiple open source tools can be used to visualize and explore Bio2RDF data (sections 3.4-3.6), as well as how dataset metrics may be used to inform querying. Future work will involve the development of a 'sandbox' for exploring and analyzing Bio2RDF data as well as the addition of more datasets through registry-compliant scripts.

5 References

1. Razick S, Magklaras G, Donaldson IM: **iRefIndex: a consolidated protein interaction database with provenance**. *BMC Bioinformatics* 2008, **9**:405.
2. Goble C, Stevens R: **State of the nation in data integration for bioinformatics**. *J Biomed Inform* 2008, **41**(5):687-693.
3. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems**. *J Biomed Inform* 2008, **41**(5):706-716.
4. Ruttenberg A, Rees JA, Samwald M, Marshall MS: **Life sciences on the Semantic Web: the Neurocommons and beyond**. *Brief Bioinform* 2009, **10**(2):193-204.
5. Momtchev V., Peychev D., Primov T., Georgiev G.: **Expanding the Pathway and Interaction Knowledge inLinked Life Data**. In: *Semantic Web Challenge: 2009; Amsterdam*; 2009.

[18] https://github.com/dbcls/bh12/wiki/Yummy-data

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

7. Campinas S PT, Ceccarelli D, Delbru R, Tummarello G: **Introducing RDF Graph Summary With Application to Assisted SPARQL Formulation.** . In: *23rd International Workshop on Database and Expert Systems Applica-tions*. Vienna Austria; 2012.

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