Bio2RDF Release 3: A Larger Connected Network of Linked Data for the Life Sciences

Michel Dumontier[1], Alison Callahan[1], Jose Cruz-Toledo[2], Peter Ansell[3], Vincent Emonet[4], François Belleau[4], Arnaud Droit[4]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, CA; [2]IO Informatics, Berkeley, CA; [3]Microsoft QUT eResearch Centre, Queensland University of Technology, Australia; [4]Department of Molecular Medicine, CHUQ Research Center, Laval University, QC

{michel.dumontier, alison.callahan, josemiguelcruztoledo, peter.ansell, vincent.emonet, francois.belleau, arnaud.droit}@gmail.com

Abstract. Bio2RDF is an open source project to generate and provide Linked Data for the Life Sciences. Here, we report on a third coordinated release of ~11 billion triples across 30 biomedical databases and datasets, representing a 10 fold increase in the number of triples since Bio2RDF Release 2 (Jan 2013). New clinically relevant datasets have been added. New features in this release include improved data quality, typing of every URI, extended dataset statistics, tighter integration, and a refactored linked data platform. Bio2RDF data is available via REST services, SPARQL endpoints, and downloadable files.

Keywords: linked open data, semantic web, RDF

1 Introduction

Bio2RDF is an open-source project to transform the vast collections of heterogeneously formatted biomedical data into Linked Data [1], [2]. GitHub-housed PHP scripts convert data (e.g. flat files, tab-delimited files, XML, JSON) into RDF using downloadable files or APIs. Bio2RDF scripts follow a basic convention to specify the syntax of HTTP identifiers for i) source-identified data items, ii) script-generated data items, and iii) vocabulary used to describe the dataset contents [1]. Bio2RDF scripts uses the Life Science Registry (http://tinyurl.com/lsregistry), a comprehensive list of over 2200 biomedical databases, datasets and terminologies, to obtain a canonical dataset name (*prefix*), which is used in the formulation of a Bio2RDF URI - http://bio2rdf.org/{prefix}:{identifier} and identifiers.org URI. Each data item is annotated with provenance, including the URL of the files from which it was generated. Bio2RDF types and relations have been mapped to the Semanticscience Integrated Ontology (SIO)[3], thereby enabling queries to be formulated using a single terminol-

ogy [4]. Bio2RDF has been used for a wide variety of biomedical research including understanding HIV-based interactions [5] and drug discovery [6].

Here, we report an update to the Bio2RDF network, termed Bio2RDF Release 3, and compare the results to Bio2RDF Release 2.

2 Bio2RDF Release 3

Bio2RDF Release 3 (July 2014) is comprised of ~11B triples across 30 datasets, 10 of which are new since Release 2 (Table 1). The top 3 datasets are Pubmed (scholarly citations; 5B triples), iProClass (database cross-references; 3B triples), and NCBI Gene (sequences and annotations; 1B triples). Compared to Release 2, there are 10x the number of triples, and each dataset has increased by an average of 300%.

Table 1. Bio2RDF Release 3 Datasets.

Dataset	new	triples	% increase	types	out links	in links
Affymetrix		86,943,196	196%	2	30	0
Biomodels		2,380,009	404%	16	50	0
BioPortal		19,920,395	125%	-	-	-
Clinicaltrials	*	8,323,598	100%	55	1	1
CTD		326,720,894	230%	9	9	1
dbSNP	*	8,801,487	100%	6	5	2
DrugBank		3,649,561	325%	69	23	2
GenAge	*	73,048	100%	2	6	0
GenDR	*	11,663	100%	4	4	0
GO Annotations		97,520,151	122%	1	6	1
HGNC		3,628,205	434%	3	11	3
Homologene		7,189,769	561%	1	4	0
InterPro		2,323,345	233%	8	18	3
iProClass		3,306,107,223	1564%	0	16	0
iRefIndex		48,781,511	157%	5	24	0
KEGG	*	50,197,150	100%	17	43	7
LSR	*	55,914	100%	1	2	0
MeSH		7,323,864	176%	7	0	4
MGD		8,206,813	334%	11	8	4
NCBI Gene		1,164,672,432	296%	16	11	11
NDC		6,033,632	34%	12	0	2
OMIM		7,980,581	432%	8	17	8
OrphaNet	*	377,947	100%	3	12	2
PharmGKB		278,049,209	733%	18	41	1
PubMed		5,005,343,905	1350%	9	0	18
SABIO-RK		2,716,421	104%	15	11	1

SGD		12,399,627	223%	42	24	1
SIDER	*	17,509,770	100%	8	3	0
NCBI Taxonomy		21,310,356	120%	5	2	12
WormBase	*	22,682,002	100%	34	5	2
Total		**10,495,601,538**		**370**	**343**	**79**

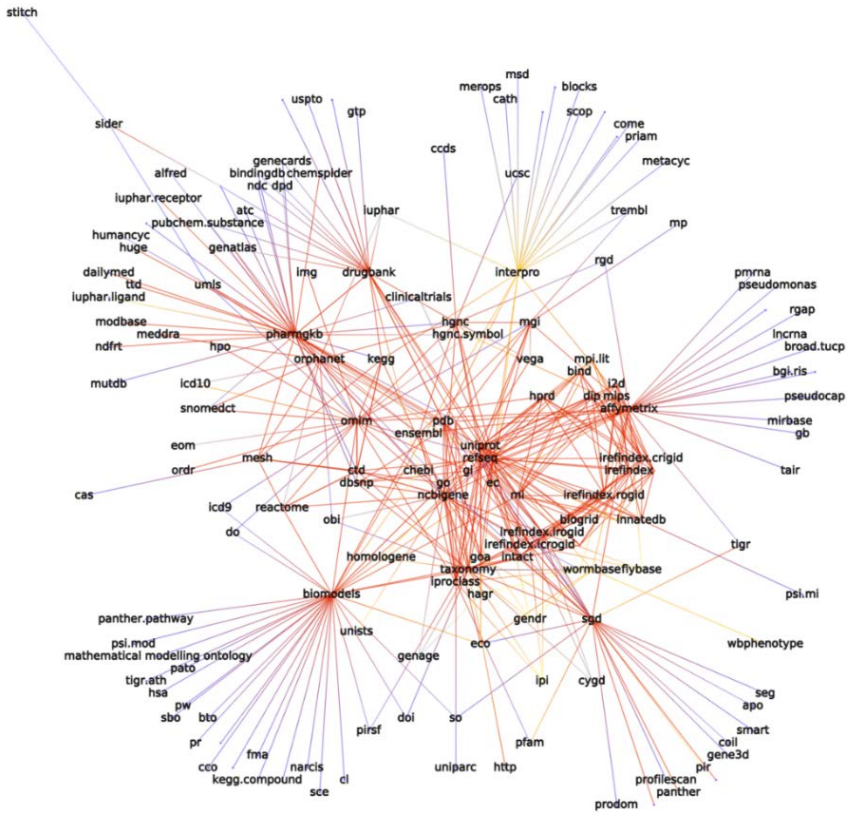

Fig. 1. Connectivity in Bio2RDF Release 3 datasets. Nodes represent datasets, edges represent connections between datasets.

Figure 1 shows a dataset network diagram using pre-computed SPARQL-based graph summaries (excluding bioportal ontologies). The network exhibits a power-law distribution, with a few highly connected nodes connected to a vast number of nodes with abut a single edge.

3 REST services

We redeveloped the Bio2RDF linked data platform to provide 3 basic services (describe, search, links) by querying the target SPARQL endpoint using Talend ESB, a graphical Java code generator based on the Eclipse framework. The REST services

now return RDF triples or quads based on content negotiation or RESTful URIs of the form http://bio2rdf.org/[prefix]/[service]/[format]/[searchterm]. The *describe* service returns statements with the searchterm as an identifier in the subject position. The *links* service returns triples with the searchterm as an identifier in the object position. Finally, the *search* service returns triples containing matched literals. Datasets and available services descriptions are stored and retrieved by the web application using a new SPARQL endpoint (http://dataset.bio2rdf.org/sparql).

4 Availability

Bio2RDF is accessible from http://bio2rdf.org. Bio2RDF scripts, mappings, and web application are available from GitHub (https://github.com/bio2rdf). A list of the datasets, detailed statistics, and downloadable content (RDF files, VoID description, statistics, virtuoso database) are available from http://download.bio2rdf.org/current/release.html . Descriptions of Bio2RDF datasets and file locations are also available from datahub.io .

5 References

[1] A. Callahan, J. Cruz-Toledo, P. Ansell, and M. Dumontier, "Bio2RDF Release 2: Improved coverage, interoperability and provenance of life science linked data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7882 LNCS, pp. 200–212.

[2] F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems," *J. Biomed. Inform.*, vol. 41, no. 5, pp. 706–716, 2008.

[3] M. Dumontier *et al*, "The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery.," *J. Biomed. Semantics*, vol. 5, p. 14, 2014.

[4] A. Callahan, J. Cruz-Toledo, and M. Dumontier, "Ontology-Based Querying with Bio2RDF's Linked Open Data.," *J. Biomed. Semantics*, vol. 4 Suppl 1, p. S1, 2013.

[5] M. A. Nolin, M. Dumontier, F. Belleau, and J. Corbeil, "Building an HIV data mashup using Bio2RDF," *Brief. Bioinform.*, vol. 13, pp. 98–106, 2012.

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