

Bio2RDF Network Of Linked Data

Marc-Alexandre Nolin ^{1,4}, Peter Ansell ², François Belleau ¹, Kingsley Idehen ³,
Philippe Rigault ¹, Nicole Tourigny ⁴, Paul Roe ², James M Hogan ², Michel
Dumontier ⁵

¹Centre de recherche du CHUL, francoisbelleau@yahoo.ca, marc-alexandre.nolin@genome.ulaval.ca, philippe.rigault@genome.ulaval.ca, ²Queensland University of Technology, p.ansell@qut.edu.au, p.roe@qut.edu.au, j.hogan@qut.edu.au, ³OpenLink Software, kidehen@openlinksw.com, ⁴Université Laval, nicole.tourigny@ift.ulaval.ca, ⁵Carleton University, michel.dumontier@carleton.ca

Abstract. Background: The Bio2RDF project (<http://bio2rdf.org>) work to create a network of coherent linked data across the life sciences databases. The project is open source and the following result is from the input of this community. Results: Databases have been converted and linked together with semantic web technologies. The process is to normalize any external URIs in each databases while we do the conversion from the original format to RDF. Conclusion: Each converted databases have his own SPARQL point provided by a Virtuoso Triplestore. Every documents can be retrieve using a REST interface similar for any databases. The REST URL also happen to be the URI. Other tools are also available using the REST interface.

Keywords: Linked data, semantic web, URI normalization

1 Introduction

In the State of the nation in data integration for bioinformatics, Goble and Stevens (2008) advised the bioinformatics community to address the issue of identity and naming, a necessary condition to facilitate data integration. As part of the Bio2RDF project to build an integrated bioinformatics warehouse on the semantic web, resources have been assigned Universal Resource Identifiers (URIs) that are normalized around the bio2rdf.org namespace (Belleau, F., et.al., 2008). Bio2RDF has created an RDF warehouse that serves over 70 million triples describing the human and mouse genomes (Belleau, F., et.al, DILS2008 in press).

The Linked Data¹ initiative aims to make possible the browsing of information on the semantic web, and follows four basic rules. First, resources should be named with Universal Resource Identifiers, or URIs. Second, HTTP URIs are ideal as they provide ownership and resolution. Third, information should be resolvable on the web, and fourth, it is necessary to connect data on the web. Building a critical mass of linked data on the web is a crucial goal for the semantic web. DBpedia (Auer, S.,

¹ <http://www.w3.org/DesignIssues/LinkedData.html>

et.al., 2007), the RDF version of Wikipedia, is fast becoming a central hub for concepts due to its multi-disciplinary nature, being connected to a large number of other public RDF sources using linked data principles. Hosted on a Virtuoso RDF server, DBpedia provides 200 million triples describing 2.18 million topics.

While traditional web technologies require HTML web browsers and text search engines to find appropriate content, semantic web browsers such as Tabulator² or, OpenLink RDF browser³ surf a web of RDF documents. Semantic search engines such as Sindice (Tummarello, G., et.al., 2008) and Swoogle (Ding, L., et.al., 2004) crawl and index RDF documents. SPARQL, a W3C recommendation⁴, is the equivalent language to SQL for querying the semantic. On a linked data semantic web, if you can browse RDF data with a semantic browser, you should also be able to query it with SPARQL.

While the previous version of the Bio2RDF project put all the data on the same server, we knew that we had to aimed to a distributed network of server to provide the RDF documents. In this papers, we will show how the 2.3 billions triples available trough our systems have been constructed, the network architecture, the software installed on every mirror and the software available publicly on Sourceforge.net. Finally, we will show an example of a query that span multiple databases using normalized URIs.

2 Methods

2.1 RDF Creation And Normalization

The Banff Manifesto⁵ asserts a best practice in creating content for the bioinformatics semantic web. Since Bio2RDF merely makes public content available, it generates normalized, producer-agnostic HTTP URI's to describe resources along with references to other databases. This makes it possible to make statements about producer content, but maintain a linked and resolvable semantic web.

The initial Bio2RDF software release brought together a number of major databases with a set of openly available JSP-based rdfizer scripts. The next version of the Bio2RDF software enhanced this process by applying semi-permanent and temporary storage of rdfized data in a SPARQL enabled RDF Database, Virtuoso. The database

² <http://www.w3.org/2005/ajar/tab>

³ <http://demo.openlinksw.com/DAV/JS/rdfbrowser/index.html>

⁴ <http://www.w3.org/TR/rdf-sparql-query/>

⁵ <http://bio2rdf.org/bm>

can be loaded, with Virtuoso's SPARQL INSERT syntax, using N3 files, available from <http://bio2rdf.org/download>, along with a bash script, `load_ttl.sh`, which can be found in the sourceforge package. The package can then respond to queries using Bio2RDF normalised URI's to determine what links exist between the different databases.

Wikipedia topics about genes, proteins, molecules are an excellent starting point in order to reference biological concepts whose definition is accepted by a majority of people. DBpedia provides permanent Semantic Web URI for these concepts. For example, the hexokinase gene is referenced by <http://dbpedia.org/resource/HK1> which in turn is represented by the Wikipedia article at <http://en.wikipedia.org/wiki/HK1>. We incorporated DBpedia's RDF triples based on Wikipedia infoboxes⁶ and loaded these into Virtuoso.

2.2 Network Architecture

The Bio2RDF network is a loosely coupled set of RDF databases which can respond to queries for the RDF versions of particular records on bioinformatics databases that they have information about. To enable this network to function reliably, some redundancy needs to be available, and queries need to be efficient. In order to provide for these two goals a combination of DNS and higher level redirection capabilities has been created to resolve queries for the normalized Bio2RDF URI's. New participants in the system do not need to necessarily provide new RDF databases, they can be simple consumers, or work as redirection points for the users inside their organization. If new participants do want to become providers, they should publish the details of where people can access their RDF database from on a community wiki, where it can be reviewed by others. If the other community participants add the SPARQL endpoint details and which databases are available on each endpoint, to their local configurations, they can test out the validity of the data. If the data provided fits with the Bio2RDF goals to have simple resolvable URI's for linked bioinformatics data, then the details of the endpoint and which databases are available will be added to the public configuration, which is currently updated with each new version of the sourceforge package. In future this configuration will be stored in RDF form, and accessible either from a local file or via a SPARQL call, thus enabling live changes to the configuration of endpoints in response to new developments.

3 Results

3.1 Bio2RDF with Virtuoso triplestore

To realize the billions triples competition, we create as much Virtuoso instance as namespaces we decide to convert. Each of these instance has been given his own sub

⁶ http://downloads.dbpedia.org/3.0/en/infobox_en.nt.bz2

domain names so the SPARQL point could be asked individually. For example, the complete Uniprot database has been converted and queryable in SPARQL at <http://uniprot.bio2rdf.org/sparql>. When a query for an URI is sent to a Bio2RDF server, a JSP script will send the right SPARQL queries to the right Virtuoso instance to fetch the document requested. Here is a list of all the database and SPARQL point currently available.

Table 1. List of databases converted, number of triples and location of the SPARQL point.

Database Name	Number of Triples	SPARQL Point
Gene Ontology	2,112,358	Http://go.bio2rdf.org/sparql
OMIM	765,384	Http://omim.bio2rdf.org/sparql
PubMed *	797,000,000	Http://pubmed.bio2rdf.org/sparql
NCBI GeneID	172,931,628	Http://geneid.bio2rdf.org/sparql
Uniprot	338,602,962	Http://uniprot.bio2rdf.org/sparql
UniRef *	242,000,000	Http://uniref.bio2rdf.org/sparql
UniParc *	490,000,000	Http://uniparc.bio2rdf.org/sparql
KEGG Pathway	84,715,161	Http://kegg.bio2rdf.org/sparql
Commons Pathway	27,623,683	Http://cpath.bio2rdf.org/sparql
Reactome	2,980,230	Http://reactome.bio2rdf.org/sparql
BioCyc	18,532,342	Http://biocyc.bio2rdf.org/sparql
MeSH	654,198	Http://mesh.bio2rdf.org/sparql
PDB (no atoms)	916,207	Http://pdb.bio2rdf.org/sparql
ChEBI	508,337	Http://chebi.bio2rdf.org/sparql
KEGG Compound	148,893	Http://kegg.bio2rdf.org/sparql
KEGG GLYCAN	94,021	Http://kegg.bio2rdf.org/sparql
Enzyme Commision	396,594	Http://ec.bio2rdf.org/sparql
KEGG Reaction	96,036	Http://kegg.bio2rdf.org/sparql
KEGG Drug	59,595	Http://kegg.bio2rdf.org/sparql
Taxonomy	3,877,201	Http://taxonomy.bio2rdf.org/sparql
PID	231,340	Http://biopax.bio2rdf.org/sparql
Dbpedia	41,606	Http://dbpedia.bio2rdf.org/sparql
HGNC	920,629	Http://hgnc.bio2rdf.org/sparql
HomoloGene	2,668,903	Http://homologene.bio2rdf.org/sparql
IproClass	149,342,977	Http://iproclass.bio2rdf.org/sparql
MGI	1,887,729	Http://mgi.bio2rdf.org/sparql
CellMap	172,638	Http://biopax.bio2rdf.org/sparql
INOH	234,493	Http://inoh.bio2rdf.org/sparql
OBO	2,145,237	Http://obo.bio2rdf.org/sparql

3.2 Full text search with semantic ranking ordering

The discovery of knowledge is facilitated by a full text search over all the RDF literals. While Virtuoso offers full text search, it does not provide a hit ranking

method. We implemented a simple ranking scheme in which full text search results are sorted according to the LinkRank statistics, which is simply the sum of inbound and outbound links (Belleau, F., et.al., DILS2008 In press). The URL syntax for this service is <http://bio2rdf.org/search/QUERY>.

By resolving <http://bio2rdf.org/search/HK1>, we initiate a full text search in the graph, then the server returns a graph containing the searched literal with matched resources. Figure 1 illustrates browsing the search results using Piggy Bank (Huynh, D., et.al., 2007), an RDF facet browser for Mozilla Firefox that filters and sorts the results and provides visualization for RDF documents. The search reveals 35 topics from 8 different data sources. The IUBMB Enzyme Nomenclature identifier [ec:2.7.1.1](http://ec2.7.1.1) obtains the highest LinkRank score of 125, indicating that it is the most referenced topic.

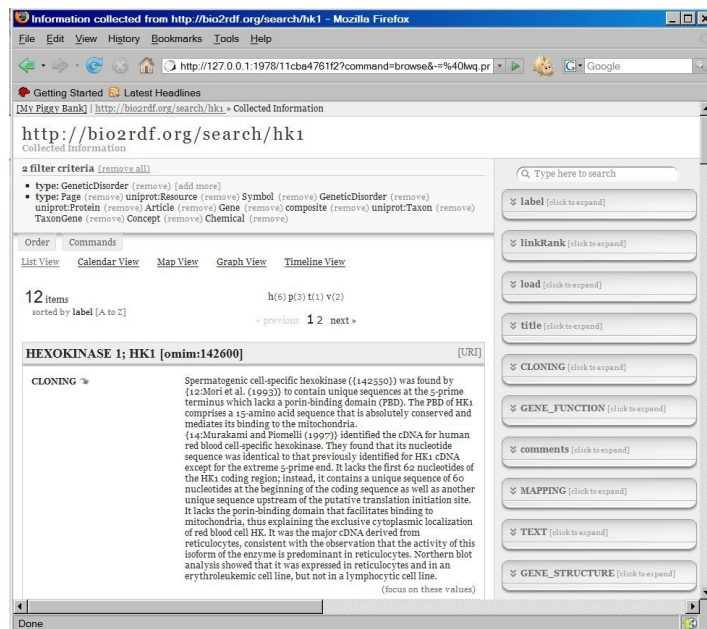


Figure 1: Example of the search utility provided by Bio2RDF view in Simile Piggy Bank

3.3 The reverse link service

Bio2RDF offers a reverse link service⁷ to determine whether a URI has been used in other sources. The reverse link service returns a set of RDF statements where the identifier is the object of the triple, hence makes it possible to discover new links to

⁷<http://bio2rdf.org/links/NAMESPACE:IDENTIFIER>

other data sources. For instance <http://bio2rdf.org/links/geneid:15275> returns uniprot:P17710 and homologene:10090-15275, hence reversibly linking these resources. Has of now, the reverse links works because we have a Virtuoso instance where many databases has been put together. But in a future version, where the data will be distributed, the reserve links won't be that easy to find. Since we know the map of what is linked to what, we may asked only the server who may know about a certain entity (Figure. 2). This procedure will reduce greatly the amount of query needed to be done to answer a reverse link request.

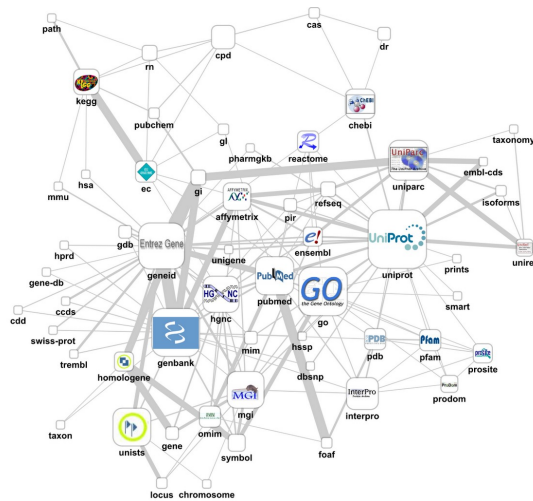


Figure 2: Map of the connexion between databases. The density of the edge is relative to the number of link between them and the size of the database is relative to the number of incoming links.

3.4 The Mirror Network

To increase the stability and reliability of the whole network, we are beginning to introduce partial mirror. A partial mirror is a server which will host some of the data of the network and not all of it. A partial mirror will redirect to another mirror which may hold the document when such document is not available on the current server. The Bio2RDF project currently have a mirror from the company Openlink Software which are the creator of the Virtuoso triple store. Other mirror project are currently on the way of being realized.

3.5 SPARQL Example

The following is a very simple example but it shows that we can recreate a kind of NCBI Entrez to search multiple database in only one query. The following query return all knowledge entity that may have the word “ HK1 “ in it and return it in number of occurrences. This query can be tried has it is directly on the main Bio2RDF.org SPARQL Point at <http://bio2rdf.org/sparql>

```
SELECT ?type1, ?label1, count(*)
WHERE {
?s1 ?p1 ?o1 .
?o1 bif:contains " HK1 " .
?s1 <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> ?
type1 .
?s1 <http://www.w3.org/2000/01/rdf-schema#label> ?
label1 .
}
ORDER BY DESC (count(*))
```

4 Discussion

A significant challenge in bioinformatics involves the integration of data originating from over 1000 biological databases (Galperin, M., 2008). A first step to this problem involves the standard assignment of identifiers for these resources, which is done in Bio2RDF using web resolvable names adhering to a consistent naming scheme. By simply pasting the name of the resource in a browser, users may now discover pertinent knowledge that has been aggregated by Bio2RDF via the open linked nature of semantic web technologies. Furthermore, in the provision of three services, we have increased the value of data represented with RDF. First, ranking of full text search results with LinkRank yields an effective method to identify those resources that are most highly linked in the known data. Second, reversible links offer the tantalizing capability of discovering links beyond the Bio2RDF store and into other triple stores. Third, a SPARQL endpoint provides the sheer flexibility to create sophisticated queries that may traverse any of the 30 data providers, including those links to linked triple stores (i.e. DBpedia).

Our role as a proxy introduces a major issue: should Bio2RDF server go down, it acts as a single point of failure and prevents access to the linked data. This is why we are going down the road of mirror network and DNS round robin balancing. The Bio2RDF mirroring process enable users not only to create their own local mirror of the official website and database, but also to stay connect with the official network of linked data. And if your lab is committed to become a data provider, follow the Banff Manifesto's rules and contact the community so we can reference you too. Talk are still in progress within the Bio2RDF mailing list about what would be the best way to create a stable, reliable and flexible network. We have some options on the

tables and we believe we will find a balance between these needs where everybody should be satisfied.

Conclusion

With the availability of billions of triples describing post genomic knowledge, linked together by Bio2RDF and other linked data sources, the development of efficient software to query this data is a major issue. While Virtuoso can be used to provide the database and query engine needed to support the overall consumption and exploration of rdfized information, we have developed new Bio2RDF services such as full text search with semantic rank and SPARQL endpoint as valuable tools for knowledge discovery. Some work are still needed regarding these tool to distributed their functionality now that we are aiming to a network of Bio2RDF servers. Other community involvement is also required to extend the Bio2RDF capabilities towards satisfying wide ranging needs, particularly in formulating more sophisticated queries in a more natural manner.

Acknowledgments

François Belleau was a recipient of a studentship from Génome Québec and Marc-Alexandre Nolin was a recipient of a studentship from the Canadian Institutes of Health Research. Peter Ansell was supported by a scholarship from the Microsoft QUT eResearch Centre. Michel Dumontier is supported by an NSERC Discovery Grant for research in semantic knowledge management.

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. *J. Mol. Biol.* 147, 195--197 (1981)
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) *Euro-Par 2006*. LNCS, vol. 4128, pp. 1148--1158. Springer, Heidelberg (2006)
3. Foster, I., Kesselman, C.: *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: *10th IEEE International Symposium on High Performance Distributed Computing*, pp. 181--184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: *The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration*. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov>