Bio2RDF: Linked Data for the Life Sciences

José Cruz-Toledo\textsuperscript{1}, Alison Callahan\textsuperscript{1} and Michel Dumontier\textsuperscript{1,2,3}

\textsuperscript{1}Department of Biology, \textsuperscript{2}Institute of Biochemistry, \textsuperscript{3}School of Computer Science Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada
{jctoledo, acallaha}@connect.carleton.ca, michel_dumontier@carleton.ca

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 describe Bio2RDF Release 2 which features updated open source scripts, improved data connectivity, provenance, statistics, CORS and SPARQL 1.1 federation friendly endpoints, and downloadable data and indexed RDF databases.

\textbf{Keywords:} Semantic Web, Linked Data, Life Sciences Data, SPARQL

1 Motivation

Any scientist who works with a variety of datasets from different sources understands the frustration of integrating multiple datasets and trying to decipher the implicit semantics of a given raw data structure. Alongside this challenge is the failure of data providers to adopt a common data format, which makes data integration a challenging, tedious, inefficient and error-prone process\cite{1,2}. Semantic Web technologies such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) provide a fundamental basis for data interoperability that builds on W3C web standards.

2 Bio2RDF Release 2

Bio2RDF is an open source project that employs Semantic Web technologies to build and provide the largest network of interoperable Linked Data for the Life Sciences. Our most recent release of Bio2RDF, Release 2, features over 1 billion triples from 19 datasets\cite{1}, including 3 datasets that themselves are comprised of hundreds of resources (\textbf{Table 1}). Updated MIT-licensed RDFization scripts are available from the Bio2RDF GitHub repository\cite{2}. Adoption of GitHub as a source code management system makes it easier for any developer to contribute new scripts and get involved in improving the quality of Bio2RDF data.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Dataset & Number of Triples \\
\hline
Dataset 1 & 1 billion \\
\hline
Dataset 2 & 100 million \\
\hline
Dataset 3 & 500 million \\
\hline
\end{tabular}
\caption{Table 1: Bio2RDF dataset sizes.}
\end{table}

\textsuperscript{1}http://download.bio2rdf.org/release/2/release.html
\textsuperscript{2}https://github.com/bio2rdf/bio2rdf-scripts
Table 1. Bio2RDF Release 2 datasets listed by name (with namespace)

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Namespace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affymetrix</td>
<td>affymetrix</td>
</tr>
<tr>
<td>Biomodels</td>
<td>biomodels</td>
</tr>
<tr>
<td>Comparative Toxicogenomics Database</td>
<td>ctd</td>
</tr>
<tr>
<td>DrugBank</td>
<td>drugbank</td>
</tr>
<tr>
<td>NCBI Gene</td>
<td>ncbigene</td>
</tr>
<tr>
<td>Gene Ontology Annotations</td>
<td>goa</td>
</tr>
<tr>
<td>HUGO Gene Nomenclature Committee</td>
<td>hgnc</td>
</tr>
<tr>
<td>Homologene</td>
<td>homologene</td>
</tr>
<tr>
<td>InterPro</td>
<td>interpro</td>
</tr>
<tr>
<td>iProClass</td>
<td>iproclass</td>
</tr>
<tr>
<td>Bio2RDF scripts</td>
<td>Bio2RDF scripts</td>
</tr>
</tbody>
</table>

Bio2RDF scripts (Fig. 1) make use of a programmatically accessible resource registry that generates Bio2RDF compliant URIs, thereby ensuring a high degree of syntactic interoperability in and across every RDFized dataset. Bio2RDF scripts generate provenance records for each dataset using the W3C vocabulary of Interlinked Datasets (VoID), the Provenance vocabulary (PROV) and Dublin Core vocabulary (DC). Each dataset includes 10 detailed metrics (e.g., total number of triples, number of records of a given type, type-type relations, etc.) that serve as an overview of the contents of the Linked Data and can be used to guide querying through our publicly available CORS and federated query capable SPARQL endpoints. The metrics are programmatically computed using SPARQL queries to the linked data generated by the dataset conversion script, and then loaded into a unique named graph at the dataset SPARQL endpoint.

In an effort to more generally expose Bio2RDF data, and to make it both human and machine-accessible, each dataset can be queried using SPARQL at its dedicated endpoint, and also browsed using the Virtuoso Faceted Browser. We also generate a summary page for each endpoint, which includes a searchable view of the dataset metrics. Finally, all generated linked data files and text indexed Virtuoso-powered triple stores are freely available for download.

![Fig. 1. Overview of Bio2RDF linked data generation and provision.](http://example.com/overview.png)

---

3 Each dataset faceted browser is available at [http://[DATASET-NAMESPACE].bio2rdf.org/fct](http://[DATASET-NAMESPACE].bio2rdf.org/fct).

4 For example, the summary page for the EBI BioModels dataset is [http://download.bio2rdf.org/release/2/biomodels/biomodels.html](http://download.bio2rdf.org/release/2/biomodels/biomodels.html). A list of all dataset summary pages is available at [http://download.bio2rdf.org/release/2/release.html](http://download.bio2rdf.org/release/2/release.html).
3 Querying Bio2RDF linked data

The basic level of syntactic interoperability across Bio2RDF datasets enables powerful federated queries, and the dataset metrics can be used to aid in the construction of such queries. For example, the EBI BioModels endpoint contains information about known computational models of biological processes that have been collected from the literature, and also include semantic annotations for biochemical reactions including Gene Ontology (GO) terms to classify them. This is confirmed by the inter- and intra-namespace metrics listed on the BioModels summary page, which shows that there are 3342 links between resources with the ‘http://bio2rdf.org/biomodels’ namespace and resources with the ‘http://bio2rdf.org/go’ namespace. While GO terms are used to annotate BioModels reactions, no additional information about the terms are available in the BioModels endpoint. The NCBO BioPortal dataset, however, includes the Gene Ontology and thus provides additional information about each GO term including its label and subclass relations to other GO terms. Because of this, it is possible to query the Bio2RDF BioModels SPARQL endpoint for biochemical reactions that are subclasses of “protein catabolic process”, the label for Gene Ontology term (GO:0030163) using the SPARQL SERVICE keyword to access data from the BioPortal endpoint:

```sparql
SELECT ?go ?label COUNT(DISTINCT ?x)
WHERE {
  {?go rdfs:label ?label .
   FILTER regex(?label, "^protein catabolic process").
   SERVICE <http://biomodels.bio2rdf.org/sparql> {
   } UNION {
     {?go rdfs:subClassOf ?tgo .
      FILTER regex(?tlabel, "^protein catabolic process").
      SERVICE <http://biomodels.bio2rdf.org/sparql> {
      }
    }
  }
}
```

5 http://biomodels.bio2rdf.org/sparql
This federated query is possible because the Gene Ontology term for “protein catabolic process” has the same URI across all datasets, enforced by the Bio2RDF dataset registry.

Given that each dataset is self-describing, further efforts are required to improve interoperability at the semantic level. One way of doing this is to map Bio2RDF dataset types and relations to a global ontology. Towards this goal, we mapped Bio2RDF vocabulary elements to the SemanticScience Integrated Ontology (SIO)\(^7\) which now renders it possible to query Bio2RDF data with SIO terms\(^3\). Other terminologies and ontologies can also be used to query Bio2RDF by using the automatically generated mappings provided by via the NCBO BioPortal\(^8\).

4 Summary

Bio2RDF Release 2 includes 19 updated datasets available for querying at dedicated SPARQL endpoints, and for searching using the Virtuoso Faceted Browser. An automatically generated human-readable summary page is also available for each dataset to facilitate informed querying of the underlying data. The continued growth of Bio2RDF depends on the participation of the scientific community that produces and consumes the source data that forms our linked data network. We welcome suggestions and comments\(^9\) to help us improve Bio2RDF.

5 References


---

7 https://code.google.com/p/semanticscience/wiki/SIO
8 http://bioportal.bioontology.org/
9 https://groups.google.com/group/bio2rdf