Bio2RDF: Towards a Mashup to build bioinformatics knowledge systems

FRANCOIS BELLEAU, MARC-ALEXANDRE NOLIN, NICOLE TOURIGNY, PHILIPPE RIGAULT, JEAN MORISSETTE
Integrating Data across web

- Two ways of looking for genomic information
  - Google It !!!
  - Specialized tools like NCBI Entrez

- What about the other databases ???
  - Every year new list of bioinformatic database is available
  - Data integration difficult by traditional data warehouses
Who bells the Cat ????

- W3C !!
  - Proposed a solution based on a series of standards
  - RDF for document and OWL for ontology
  - RDF and OWL generate a triple – subject, predicate and object
  - Database systems capable of handling triples are known as triplestore
Bio2RDF – A Mashup

- Integrates data from more than one source
- Integrates data from popular public databases
- Bio2RDF is a semantic web approach for data integration
Integration using Semantic approach

- Describing and building knowledge systems for semantic web is a challenge for bioinformatic community
- A few specialized projects like YeastHub and FungalWeb are successful to a certain extent
- Bio2RDF is an attempt in this area to integrate data from different sources
Materials and Methods

- Two main ideas of development
  - Conversion of existing databases into RDF format
  - Use semantic web software to merge, query and visualize data
  - Protégé ontology editor, Piggy Bank, Welkin and LSID browser
Ontology Design

- Ontology by definition is explicit specification of conceptualization
- Analyze existing HTML pages, identify predicates and relations describing the entities
- A hyperlink corresponds to a URI and a label to its predicate
- OWL description for each selected HTML document created.
RDFizing

- RDFizer were necessary for two key objectives
  - Mapping between data elements of the original document and the predicates in RDF version
  - Normalization of URI according to Bio2RDF syntax
- RDFizer programs for Bio2RDF written in JSP
- Three kinds of RDFizing carried out
  - XML to RDF
  - SQL to RDF
  - Text to RDF
URI Normalization

- Normalized URIs needed to allow proper connection of triples
- No links would be created if more than one way of expressing URI existed.
  - http://www.geneontology.org/go#GO:0004396
  - http://purl.uniprot.org/go/0004396
  - All the above represent the same hexokinase, but they are not linked since their URIs are different
A solution in Bio2RDF

- The Strategy
  - Use a REST like interface
  - Lowercase all the URI up to the colon
  - All URIs should return an RDF document

- Syntax of a Normalized Bio2RDF URI
  - http://bio2rdf.org/<namespace>:<identifier>
- Representational State Transfer enables us to produce a stable and clear URI for every document.
- The URI case sensitivity poses a problem because each different case results in a theoretically different URI.
- If URI for a document creates web page instead of RDF, Linking of data difficult.
Bio2RDF Architecture

External data sources
- Entrez Gene
- OMIM
- GO
- PDB
- MGI
- OBO
- UniProt
- UniRef
- UniParc
- RefSeq
- Nucleotide
- JCVI Protein
- PubMed
- NEWT
- PubChem
- proSite

Bulk load XML, RDF, Text

Hosted data

URL rewrite

RDF, XML

JSP RDFizer

Triplestore

Graphic user interfaces
- Sesame
- Spreadsheets
- Piggy Bank
- Tabulator
- Isid Browser
ELMO Crawler and SESAME Interface

- Elmo crawls RDF documents from the Bio2RDF website
- Sesame interface allows users to browse and query the knowledge base with SeRQL
Three Specific Services added to allow ELMO crawl Specific Knowledge

- To obtain a list of URIs corresponding to the results of a text search using the search engine of the corresponding website.
- To request all URIs in the triplestore which belongs to the specified namespace.
- To create a synonym node to link two URIs which have the same id but different synonymous namespaces.
## Results of Bio2RDF

<table>
<thead>
<tr>
<th>Data source</th>
<th>Short URI example</th>
<th>Number of RDF documents</th>
<th>Format of source data</th>
<th>Hosted version</th>
</tr>
</thead>
<tbody>
<tr>
<td>genenames.org</td>
<td>hgnc:4922</td>
<td>27,634</td>
<td>Tabulated</td>
<td>December 2007</td>
</tr>
<tr>
<td>informatics.jax.org</td>
<td>mgi:96103</td>
<td>70,172</td>
<td>Tabulated</td>
<td>June 2007, MGI 3.54 release</td>
</tr>
<tr>
<td>genome.ad.jp</td>
<td>path:mmu00010</td>
<td>68,307</td>
<td>KGML</td>
<td>December 2007, Release 44.0+/12-19</td>
</tr>
<tr>
<td>genome.ad.jp</td>
<td>cpd:C00011</td>
<td>15,006</td>
<td>Text</td>
<td>December 2007, Release 44.0+/12-19</td>
</tr>
<tr>
<td>genome.ad.jp</td>
<td>dm:D00001</td>
<td>6755</td>
<td>Text</td>
<td>December 2007, Release 44.0+/12-19</td>
</tr>
<tr>
<td>genome.ad.jp</td>
<td>ec:2.7.1.1</td>
<td>4,958</td>
<td>Text</td>
<td>December 2007, Release 44.0+/12-19</td>
</tr>
<tr>
<td>genome.ad.jp</td>
<td>gt:G00001</td>
<td>10,972</td>
<td>Text</td>
<td>December 2007, Release 44.0+/12-19</td>
</tr>
<tr>
<td>genome.ad.jp</td>
<td>rn:R00014</td>
<td>7422</td>
<td>Tabulated</td>
<td>December 2007, Release 44.0+/12-19</td>
</tr>
<tr>
<td>ebi.ac.uk</td>
<td>chebi:16526</td>
<td>13,360</td>
<td>XML</td>
<td>December 2007, Release 39.0</td>
</tr>
<tr>
<td>rcsb.org</td>
<td>pdb:1HKC</td>
<td>48,091</td>
<td>OBO/RDF</td>
<td>December 2007</td>
</tr>
<tr>
<td>geneontology.org</td>
<td>go:0004396</td>
<td>24,634</td>
<td>RDF</td>
<td>December 2007</td>
</tr>
<tr>
<td>obofoundry.org</td>
<td>obo's 54 namespaces</td>
<td>108,955</td>
<td>RDF/RDF</td>
<td>December 2007</td>
</tr>
<tr>
<td>beta.uniprot.org</td>
<td>uniparc:UPI00005AC213</td>
<td>30,261,843</td>
<td>RDF</td>
<td>December 2007</td>
</tr>
<tr>
<td>beta.uniprot.org</td>
<td>uniprot:P19367</td>
<td>4,177,176</td>
<td>RDF</td>
<td>December 2007</td>
</tr>
<tr>
<td>beta.uniprot.org</td>
<td>uniref:Uniref50_P19367</td>
<td>7,990,452</td>
<td>RDF</td>
<td>December 2007</td>
</tr>
<tr>
<td>beta.uniprot.org</td>
<td>taxon:9606</td>
<td>441,422</td>
<td>RDF</td>
<td>December 2007</td>
</tr>
<tr>
<td>ncbi.nlm.nih.gov</td>
<td>pubmed:3207429</td>
<td>17,000,000</td>
<td>XML</td>
<td>December 2007</td>
</tr>
<tr>
<td>ncbi.nlm.nih.gov</td>
<td>pubchem:3313</td>
<td>38,000,000</td>
<td>XML</td>
<td>December 2007</td>
</tr>
<tr>
<td>reactome.org</td>
<td>reactome:70326</td>
<td>8,332</td>
<td>BioPAX/RDF</td>
<td></td>
</tr>
<tr>
<td>expasy.org</td>
<td>prosite:PS00378</td>
<td>2,819</td>
<td>HTML</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>162,778,598</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Parkinson’s Use Case

- An intro to Parkinson’s – A slow progressive neurodegenerative disorder
- Four genes Rxr, Nurr1, Nur77 and Nor-1 are of interest in parkinson’s
- Major questions that can be answered by Bio2RDF
  - Which GO terms describe our four genes of interest (Rxr, Nurr1, Nur77, and Nor-1)?
  - Which articles mentioning our four genes of interest are related to apoptosis AND cytoplasm and also mention genes having GO annotations about apoptosis OR cytoplasm?
A Simple query to find the GO terms!!

```
01 SELECT DISTINCT
02 searchLabel, geneLabel, goLabel
03 FROM
04 {search} rdf:type {<http://bio2rdf.org/bio2rdf#Search>};
05 <http://bio2rdf.org/bio2rdf#query> {searchLabel};
06 rdfs:seeAlso {gene},
07 {gene} rdfs:label {geneLabel};
08 <http://bio2rdf.org/bio2rdf#xGO> {go},
09 {go} rdfs:label {goLabel}
```
Query to find annotations of cytoplasm and apoptosis

```sql
SELECT DISTINCT
    geneSearch, geneLabel, goLabel, articleLabel
FROM
    {search} rdf:type {<http://bio2rdf.org/bio2rdf#Search>};
    <http://bio2rdf.org/bio2rdf#query> {geneSearch};
    rdfs:seeAlso {gene},
    {gene} rdfs:label {geneLabel};
    <http://bio2rdf.org/bio2rdf#xArticle> {article};
    <http://bio2rdf.org/bio2rdf#xGO> {go},
    {go} rdfs:label {goLabel},
    {article} rdfs:label {articleLabel};
WHERE
    p {literal}
    (  
        go = <http://bio2rdf.org/go:0006915>
        OR
        go = <http://bio2rdf.org/go:0005737>
    )
    AND
    (  
        literal like "*apoptosis*" ignore case
        AND
        literal like "*cytoplasm*" ignore case
    )
```
Compatibility with ongoing semantic web projects

- Bio2RDF compatible with ongoing semantic web projects
- Compatible with tabulator and various LSID browsers
- The RDF graph returned by Bio2RDF makes it compatible with facet browsers like piggy bank
Extendability and Scalability

- Simple steps to add new database sources
  - Design RDF document to represent data
  - Write corresponding rdfizer programs
  - Install new rdfizer under Bio2RDF servlet of the myBio2RDF installation
  - Add a rewrite rule to the urlrewrite.xml configuration file
  - Restart the myBio2RDF servlet
A Work in Progress..

- The ontology and rdfizer are not definitive
- The ontology still in early stages of development
- The project is open source and can be accessed at bio2rdf.sourceforge.net
THANK YOU