Linked data at EMBL-EBI

The EBI RDF Platform

Andy Jenkinson
andy.jenkinson@ebi.ac.uk
MISSION

• Data services
• Research
• Training
• Supporting Industry
• Coordinating biology data provision in Europe
What is EMBL-EBI?

• Part of the European Molecular Biology Laboratory
• Inter-governmental, non-profit research institute
• Funded by member states
• Europe’s hub for biological data services and research
• 500 members of staff from 53 nations.
Overview

• Background: what’s our motivation?
• Objectives: what we set out to do
• Services: what we have produced
• Details: how it works
• Timeline: what we will be doing from now
• Questions and answers*

*answers subject to availability
Life science: many data types

Genes, genomes & variation

Gene, protein & metabolite expression

Protein sequences, families & motifs

Macromolecular structures

Chemogenomics & metabolomics

Interactions, reactions & pathways

Cross-domain tools & resources
The facts

• Need for integrating data is increasing

• Integrating biology data is hard

• Data integration *per se* is not science
Programmatic Access to Data

Data Integration Technology Progression

- We tackle data integration at various layers
  - e.g. web services
- RDF abstracts out the syntactic layer
  - Leaving “only” the semantics
- Biological information seems ideal candidate
RDF: the best bits

• You can say anything in a series of triples
  • The “future use case”
• URIs to identify things
  • Globally unique
  • “Web native”
• Standard formats, query languages
  • Separate datasets are inherently compatible
• Concentrate only on the semantics
  • Your DB schema is your data
  • Encourages you to think about the data
Background
Background

“If you build it, he will come”
Objectives

- Why hasn’t RDF taken over biology?
  - Opaque technology
  - Lots to learn
  - Still developing, rapidly changing
  - Legacy of “toy” use cases

“0 results”

Not Found

The requested URL /foo was not found on this server.
Target audience: start modest

For:
- Software developers
- Technically-minded informaticians
  - Interested in RDF but with no experience

Not For:
- Biologists
Objectives

1. Coordinate existing activities at EBI
   - Some existing work
   - Opportunity to produce a more coherent set of resources
   - Some consistency of style where possible
   - Better interlinking between datasets
   - Stable services
Objectives

2. Pump-prime infrastructure
   • Make it easier for more groups to provide RDF:
     • Additional EBI data resources
     • Wider European infrastructure (BioMedBridges)
   • Triple store testing
   • Production-quality hardware and architecture
   • Policies and guidelines
Objectives

3. Lower barriers to entry:
   • Documentation
   • Examples
   • Centralised resources
   • Make URIs resolve
   • Some basic prettifying
Technical advisory input

- UCB/Celltech
- Syngenta
- Swiss Institute of Bioinformatics
- BBC
Basic components

1. Bulk downloads
   - Manually created RDF transformations
   - Full datasets
   - Part of production cycles

2. Public SPARQL access

3. Linked data browser
   - Human readable frontend for RDF data

4. RDF Platform website
   - Documentation
   - Consistency of style
RDF Platform

https://www.ebi.ac.uk/rdf/
The EBI RDF Platform aims to bring together the efforts of a number of EMBL-EBI resources that provide access to their data using Semantic Web technologies. It provides a unified way to query across resources using the W3C SPARQL query language. We welcome comments or questions via our feedback form.

Current RDF resources

<table>
<thead>
<tr>
<th>Services</th>
<th>Quick links</th>
<th>Example query</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioModels</td>
<td>Service description</td>
<td>All model elements with annotations to acetylcholine-gated channel complex (GO:0005892)</td>
</tr>
<tr>
<td></td>
<td>SPARQL endpoint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RDF download</td>
<td></td>
</tr>
<tr>
<td>BioSamples</td>
<td>Service description</td>
<td>Samples treated with monocyclic heteroarene</td>
</tr>
<tr>
<td></td>
<td>SPARQL endpoint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RDF download</td>
<td></td>
</tr>
<tr>
<td>ChEMBL</td>
<td>Service description</td>
<td>Find drug-like (but currently not approved) molecules which bind 7TM1 GPCRs with high affinity</td>
</tr>
<tr>
<td></td>
<td>SPARQL endpoint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RDF download</td>
<td></td>
</tr>
<tr>
<td>Expression</td>
<td>Service description</td>
<td>Under what experimental conditions is Fosamis bioactive?</td>
</tr>
<tr>
<td></td>
<td>SPARQL endpoint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RDF download</td>
<td></td>
</tr>
</tbody>
</table>
## Summary of Services

This table provides a summary of the available datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Version No.</th>
<th>Date Updated</th>
<th>No. Triples</th>
<th>Dataset description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioSamples</td>
<td>v20130809</td>
<td>Aug 9 2013</td>
<td>41,840,881</td>
<td><a href="http://rdf.ebi.ac.uk/dataset/biosamples/v20130809">http://rdf.ebi.ac.uk/dataset/biosamples/v20130809</a></td>
</tr>
<tr>
<td>ChEMBL</td>
<td>17.0</td>
<td>Aug 29 2013</td>
<td>374,762,364</td>
<td><a href="http://rdf.ebi.ac.uk/dataset/chembl/17.0">http://rdf.ebi.ac.uk/dataset/chembl/17.0</a></td>
</tr>
<tr>
<td>Reactome</td>
<td>45</td>
<td>Sep 05 2013</td>
<td>12,206,321</td>
<td><a href="http://rdf.ebi.ac.uk/dataset/reactome/45">http://rdf.ebi.ac.uk/dataset/reactome/45</a></td>
</tr>
</tbody>
</table>

These are all complete versions of their respective parent databases.
Gene Expression Atlas

The Functional Genomics Production Team are pleased to announce the beta publication of gene expression data from the Gene Expression Atlas as RDF Linked Open Data. The Expression Atlas Linked Dataset is an alternative API to the Gene Expression Atlas data. The purpose of this API is to enable richer queries over the data, it also supports federated queries over other linked datasets, including ChEMBL, Reactome, BioModels, BioSample, Uniprot, BioPortal and Bio2RDF.

The primary interface to the Expression Atlas RDF data is via the SPARQL endpoint.

Full VOID dataset description at http://rdf.ebi.ac.uk/dataset/atlas/13.07

<table>
<thead>
<tr>
<th>Title</th>
<th>Gene Expression Atlas RDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>RDF representation of all the experiments loaded into the Gene Expression Atlas database.</td>
</tr>
<tr>
<td>Version</td>
<td>13.07</td>
</tr>
<tr>
<td>Issued</td>
<td>July 18 2013</td>
</tr>
<tr>
<td>Number of triples</td>
<td>447149547</td>
</tr>
</tbody>
</table>
Expression Atlas SPARQL Endpoint

Enter SPARQL Query

```sparql
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>
PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
PREFIX atlasters: <http://rdf.ebi.ac.uk/terms/atlas/>

SELECT DISTINCT ?experiment ?description WHERE {
  ?experiment a atlasters:Experiment .
  ?experiment dcterms:description ?description
}
```

Example Queries

- **Query 1**
  Get experiments where the sample description contains diabetes

- **Query 2**
  Get differentially expressed genes where factor is asthma

- **Query 3**
  Show expression for ENSG00000129991 (TNNI3)

- **Query 4**
  Show expression for ENSG00000129991 (TNNI3) with its GO annotations from Uniprot (Federated query to http://beta.sparql.uniprot.org/sparql)

- **Query 5**
  For the genes differentially expressed in asthma, get the gene products associated to a Reactome pathway

- **Query 6**
  Get all mannins for a given probe e.g. AFFY-1/661 at
Interlinking datasets

For a given compound, find:

• Pathways that include the protein it targets
• Factors for which genes are differentially expressed
Tech Stuff
Content Negotiation

Linked Data browser:

```
curl -L -H "Accept: text/html" \\  
http://rdf.ebi.ac.uk/resource/atlas/E-GEOD-14539
```

RDF:

```
curl -L -H "Accept: application/rdf+xml" \\  
http://rdf.ebi.ac.uk/resource/atlas/E-GEOD-14539
```

RDF (VoID/prov):

```
curl -L -H "Accept: application/rdf+xml" \\  
http://rdf.ebi.ac.uk/dataset/atlas/13.07
```
URI Policies

• All “EBI-owned” RDF data will get URIs under:
  • http://rdf.ebi.ac.uk
    e.g. http://rdf.ebi.ac.uk/resource/atlas/E-GEOD-14539

• IDs must be registered in identifiers.org

• Collaborations normally have their own domains, and are free to choose
  • Could choose Identifiers.org as source of URIs:
    e.g. http://identifiers.org/reactome/REACT_238.2

• When linking, try to ask data owners for their preferences
  • persistent, unique, de-referencable
    e.g. http://purl.obolibrary.org/obo/GO:0005886
First release (September 2013)

- Six datasets
- Mostly production services
  - UniProt SPARQL endpoint updating soon
- Mostly stable data, but some changes to be expected
  - e.g. W3C provenance not final
  - BioSamples liable to change
- Real-world usage testing of SPARQL endpoints
- Act on feedback
From Now

- Data relatively stable
- Ready for integration into downstream applications
- SPARQL architecture may require changes
- Collecting feedback, e.g.:
  - Data models
  - Desirable data (no promises!)
  - Feature requests for browser, etc.
- Other EBI resources may choose to begin process of providing RDF
Q3 2015

- Project review
- Look at usage statistics
- Assess penetration, utility, take-up
- Range of possible outcomes
  - Roll out across all data resources
  - Maintain
  - If nobody uses it: discontinue.
Acknowledgements

• Simon Jupp, James Malone
• Mark Davies, Anna Gaulton
• Sarala Wimalaratne
• Marco Brandizi
• Leyla Garcia, Jerven Bolleman
• Joseph Rossetto, Dietmar Sturmayr
• Ewan Birney
Funding

Components of the RDF platform pilot are supported by a number of sources, including:

• EMBL

• European Commission:
  • BioMedBridges [284209]
  • Diachron [601043]

• OpenPhacts (Innovative Medicines Initiative)

• National Institutes of Health
Questions?

Sign up for our mailing list:

http://listserver.ebi.ac.uk/mailman/listinfo/rdf-announce