Introduction to RDF and the Semantic Web for the life sciences

Simon Jupp
Sample Phenotypes and Ontologies Team
European Bioinformatics Institute
jupp@ebi.ac.uk
Outline

• Overview of RDF and Semantic Web
• Hands on practical
  • Modeling data in RDF
  • Working with RDF triple stores
  • Querying data with SPARQL
• EBI RDF platform
• Use cases – ChEMBL RDF
Learning outcomes

• What is RDF and how to use it
• How to model data in RDF
• What is an Ontology
• How to use an RDF database (triple store)
• How to query RDF data with SPARQL
• How to use the EBI RDF platform
Disclaimers

- Steep learning curve
  - Lots of new concepts to learn in short time
- Hopefully a practical introduction to this technology
  - There’s a lot that we won’t cover
- All new practical session
  - Gives us feedback, ask questions.
- This technology is still new and sometimes unforgiving
  - Technology maturing
  - Best practices still evolving
Some biological questions

“Differentially expressed genes in adult mice, bred in oxygen rich vs oxygen poor environments? Of this set, which biological processes (GO) are enriched?”

“Where are genes with antigen binding function differentially expressed, which disease and which associated pathways?”

“Get metformin associated pathways with differentially expressed genes, find any proteins that are targets for known diabetes drugs”

How do you go about answering these kinds of questions?...
...you look for the data

Genes, genomes & variation
- European Nucleotide Archive
- Ensembl
- Ensembl Genomes
- European Genome-phenome Archive
- Metagenomics portal

Gene, protein & metabolite expression
- ArrayExpress
- Expression Atlas
- Metabolights
- PRIDE

Literature & ontologies
- Europe PubMed Central
- Gene Ontology
- Experimental Factor Ontology

Protein sequences, families & motifs
- InterPro
- Pfam
- UniProt

Molecular structures
- Protein Data Bank in Europe
- Electron Microscopy Data Bank

Chemical biology
- ChEMBL
- ChEBI

Reactions, interactions & pathways
- IntAct
- Reactome
- MetaboLights

Systems
- BioModels
- Enzyme Portal
- BioSamples
Doing science from data

Biological question

Where are the answers?

Ensembl  Reactome  UniProt  ChEMBL

Write parsers, mappings, code, applications, analysis...

How much does this cost?
• Questions change
• Databases change
• New databases emerge
• Technologies change

New biological question.. repeat
Problems are rooted in the data

• Heterogeneous formats and identifiers
• We invest heavily in mapping and cross-linking resources, but it’s still hard to integrate and query across internal/external resources.
  • Lots of effort doing mapping, each groups duplicate these efforts
How we deal with it in practice

• Army of Bioinformaticians
  • Coders (Perl, Python, Java, scripts)

• Develop APIs and Web services
  • Pros: Solution like SOAP/XML and REST/JSON etc provided unified access to data. Protect from underlying data model changes
  • Cons: Restrict access to underlying data. No semantics, hard to integrate data from different sources

• Warehousing / centralised services
  • Pros: Data integrated up front, good for fast queries
  • Cons: High maintenance, managing data concurrency. What about when you want to integrate with data outside of the warehouse?
NoSQL and BigData™ revolution

• Relational data, still the norm and will be for most situations
• Ways of thinking about semi-structured, complex and big data are changing
  • Document databases (mongodb)
  • Key-value stores: schema-less, distributed, big data analysis (map-reduce)
  • Graph databases: schema and data intertwined, natural way to think about data, good for complexity (biology)
Graph databases

- Graphs are “whiteboard friendly”, the reflect how you would naturally think about the data.

- The data is schema-less, or the data is the schema.

- Easy to extend and expand

- They are very good for complex data.

- Popular graph database: Neo4j

http://www.cbsolution.net/techniques/ontarget/databases_relational_vs_object_vs
Take home number 1

• Resource Description Framework (RDF) provides a data model for representing graph data structures
  • RDF is a standard vocabulary for how you describe “resources” in terms of relations to other “resources”

• RDF is built and designed for the Web
Evolution of the web

- 1\(^{st}\) generation web - linked documents
  - Hand coded HTML
- 2\(^{nd}\) generation web – the web as platform
  - APIs, Web services, XML and JSON
- 3\(^{rd}\) generation web – The web of **things** (or Semantic Web)
  - The Web as a platform for publishing data (in RDF)
  - Semantic markup that adds meaning to data for machine processing
The Web at 25?

A brief history...

- 1995-2004 W3C develop specification for a vocabulary for Web meta-data called Resource Description Framework (RDF)

http://en.wikipedia.org/wiki/Barack_Obama

Publishing data as a graph

http://en.wikipedia.org/wiki/Barack_Obama

Person

President of the United States

dbpedia:Barack_Obama

“05:23, 18 March 2004”

type

created

about

birthdate

birthplace

Honolulu

type

1961-08-04

EMBL-EBI
A brief history…

• W3C develop vocabulary for describing RDF schemas (RDFS)
  • The introduction of the rdfs:subClassOf (is-a) relation
  • A transitive property
  • Arguably the most useful relation in bio-medical ontology
The Semantic Web

- Standardise the syntax for representing data
- Graph based datamodel, built from RDF statements (triples)
- Use existing Web technology to provide global identifiers for things using Uniform Resource Identifiers (URIs)
  - e.g. http://mydomain/mythings/myhat
- Built on a technology W3C (Standards)
  - Vocabularies with strict semantics – so we can talk about “things” and the relations between “things” that a computer can interpret
  - Infer new relationships between things
Also known as Linked Data

"The Semantic Web is a **webby** way to link data"

“The existing web links **documents**, the semantic web links **data**”

“Turning the web into a **global** API”

“Shared meaning through **ontologies**”
Linked data principles

★ make your stuff available on the Web (whatever format) under an open license

★★ make it available as structured data (e.g., Excel instead of image scan of a table)

★★★ use non-proprietary formats (e.g., CSV instead of Excel)

★★★★ use URIs to identify things, so that people can point at your stuff

★★★★★ link your data to other data to provide context

Life sciences leading the way here
A resource of information

http://en.wikipedia.org/wiki/Barack_Obama

Lots of data hidden in this page

How could a machine pull data about US presidents from Illinois?
Data as a graph

- We can draw the bits of data in a graph
RDF – Resource Description Framework

• RDF is a graphical language used for representing information about resources on the web.

• Resources are described in terms of properties and property values using RDF statements.

• All statements in RDF are triple, consisting of a subject, predicate and object.
Anatomy of a triple statement

- Subject, predicates and objects (and nothing else!)

Subject: Barack Obama
Predicate: birth place
Object: Honolulu
Identify things on the web

- Using existing Web technology – global identifiers using the URI

http://dbpedia.org/page/Barack_Obama

Subject

http://dbpedia.org/property/birthPlace

Predicate

http://dbpedia.org/page/Honolulu

Object
Serialisation of RDF

- Lots of RDF serialisations
  - RDF/XML, Turtle, N-Triples, JSON-LD
  - Serialisation is for the computer

```xml
<rdf:RDF>
  <rdf:Description rdf:about="http://dbpedia.org/resource/Barack_Obama">
    <dbpedia-owl:birthPlace rdf:resource="http://dbpedia.org/resource/Honolulu"/>
    <foaf:givenName xml:lang="en">Barack</foaf:givenName>
  </rdf:Description>
</rdf:RDF>
```
N-triples a basic triple representation


Turtle is an abbreviated form of N-triples

PREFIX dbpedia:<http://dbpedia.org/page/>
PREFIX dbproperty:<http://dbpedia.org/property/>
PREFIX foaf:<http://xmlns.com/foaf/>

dbpedia:Barack_Obama dbproperty:birthPlace dbpedia:.Honolulu
dbpedia:Barack_Obama foaf:name “Barack Obama”

Note:
Objects of a triple can be a resource or a value (string, double, float)
Types of things

Subject: http://www.ebi.ac.uk/~jupp
Predicate: rdf:type
Object: Human

Subject: http://www.ebi.ac.uk/~jupp
Predicate: rdf:type
Object: Bioinformatician

Subject: http://www.ebi.ac.uk/~jupp
Predicate: rdf:type
Object: Male
Types of things

We describe these “Things” using ontologies

Subject                        Predicate                      Object
http://www.ebi.ac.uk/~jupp     rdf:type                       http://www.ebi.ac.uk/~jupp
                                 rdf:type                       Human
                                 rdf:type                       Bioinformatician
                                 rdf:type                       Male
What are Ontologies?

Different Words Same Concept

Dog

Canine
The Ogden Triangle

- Humans require words (or at least symbols) to communicate efficiently. The mapping of words to things is only indirectly possible. We do it by creating concepts that refer to things.
- The relation between symbols and things has been described in the form of the meaning triangle:

“Roast Beef”

[Ogden, Richards, 1923]
We put things into categories

- All these instances hang about making our world
- Putting these things into categories is a fundamental part of human cognition
- Psychologists study this as *concept formation*
- The same instances are put into a category
We need to know what we’re talking about…

• … if we don’t, our data are useless
• If we are to interpret our data then we need to know what entities it describes
• We need to share data and re-use it
• We need to find data; compare data; analyse data
• We need to know what we know…. 
Web Ontology Language – (OWL)

- W3C standard vocabulary for describing ontologies
  - We can describe sets of things in OWL
    - e.g. SubClassOfAxiom (A,B) - Implies all things of type A, are also things of type B
  - Powerful knowledge representation language that can be expressed in RDF

mitochondrial chromosome ‘equivalent to’ chromosome and ‘part of’ some mitochondrion
Ontologies for life sciences

Genotype
- Sequence
  - Gene products
  - Transcript
- Gene products
- Transcript

Proteins
- Sequence types and features
- Genetic Context
- Protein covalent bond
- Protein domain
- UniProt taxonomy

Pathways
- Pathway ontology
- Event (INOH pathway ontology)
- Systems Biology
- Protein-protein interaction

BRENDA tissue / enzyme source

Anatomy
- Mosquito gross anatomy
- Mouse adult gross anatomy
- Mouse gross anatomy and development
- C. elegans gross anatomy
- Arabidopsis gross anatomy
- Cereal plant gross anatomy
- Drosophila gross anatomy
- Dictyostelium discoideum anatomy
- Fungal gross anatomy FAO
- Plant structure
- Maize gross anatomy
- Medaka fish anatomy and development
- Zebrafish anatomy and development

Development
- Arabidopsis development
- Cereal plant development
- Plant growth and developmental stage
- C. elegans development
- Drosophila development FBdv fly development.obo OBO yes yes
- Human developmental anatomy, abstract version
- Human developmental anatomy, timed version

Phenotype
- NCI Thesaurus
- Mouse pathology
- Human disease
- Cereal plant trait
- PATO PATO attribute and value.obo
- Mammalian phenotype
- Habronattus courtship
- Loggerhead nesting
- Animal natural history and life history

eVOC (Expressed Sequence Annotation for Humans)
Relational Data to RDF graph conversion

- Give “things” URIs
- Type “things” with ontologies
- Link “things” to other related “things”
- We can do this with any kind of data
Atlas RDF Schema
http://rdf.ebi.ac.uk/terms/atlas

Gene Expression Atlas RDF schema

Experiment
- pubmed
- samples
  - ex factor
  - Diff expression
  - ENSEMBL genes

ENSEMBL genes
- UniProt
- Reactome
- ChEMBL
Strings to “things”

Moving from URLs and strings to URLs

<table>
<thead>
<tr>
<th>Human readable</th>
<th>Human or machine readable</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.ebi.ac.uk/gxa/experiment/E-MEXP-3392">http://www.ebi.ac.uk/gxa/experiment/E-MEXP-3392</a></td>
<td>ENSRNOG00000002940</td>
</tr>
<tr>
<td><a href="http://rdf.ebi.ac.uk/resource/atlas/E-GEOD-1005">http://rdf.ebi.ac.uk/resource/atlas/E-GEOD-1005</a></td>
<td><a href="http://identifiers.org/ensembl/ENSRNOG00000002940">http://identifiers.org/ensembl/ENSRNOG00000002940</a></td>
</tr>
<tr>
<td>‘Compound treatment’ ‘endothelin-1 100 nanomolar’</td>
<td></td>
</tr>
</tbody>
</table>
XML/JSON view

Just an exchange format
no shared meaning and no explicit links
RDF/XML view “Self documenting”

Resource and links are explicit (URIs)
Shared meaning/semantics (Ontologies)
http://purl.uniprot.org/uniprot/Q5UAB1
Now serving up RDF

http://purl.uniprot.org/core/Q5UA70
Storing and querying RDF

- Optimized databases for RDF data
- SPARQL query language

Stardog

Apache Jena

OWLIM

Virtuoso

openRDF.org

Embl-EBI
Querying the data with SPARQL
SPARQL

- W3C standard query language for querying RDF data
  - Query language for matching graph patterns in RDF
  - SQL like: SELECT, DISTINCT, COUNT, WHERE, GROUP, ORDER, LIMIT, OFFSET …

```sql
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?subject WHERE {
  ?subject rdf:type umbel:PresidentOfOrganization .
}
```
Recap

- RDF is a data model for describing Graphs
  - Based on simple triple format
- Well suited for complex data and supporting data integration
- Describe what the data is using controlled vocabularies and ontologies
  - Add semantics to infer new relationships
- Built for the Web on existing Web technologies
- SPARQL is the RDF query language