Swobjects: SPARQLing away at data integration – drinking your data from the source

Helena F. Deus, Scott Marshall and Eric Prud’hommeaux
Digital Enterprise Research Institute
What is the role for SW in LF?
Biology is becoming an information science!

- Cell and computers are not that different

\[
\frac{dV_{\text{out}}}{dt} + \sigma(t)V_{\text{out}} = f(t)
\]

\[
V_{\text{sample}} = \int_{\tau} \left( V_{\text{in}}(\tau) \hat{h}(\tau) + h_{1}(\tau) \right) d\tau
\]

Activator genes

Repressor genes
Linked data for Life Sciences

RAW DATA!

Interpretation of Results!

Enabling networked knowledge.
Linked data, Networked Knowledge

Network of Knowledge!
Is Linked Data Semantic Web in new clothes?

“Linked Data refers to a set of best practices for publishing and connecting structured data on the Web.”
- Chris Bizer, 2009

“In the Semantic Web, it is not the Semantic which is new, it is the Web which is new”.
- Chis Welty/IBM

We need to “put the Web back in Semantic Web”
- ISWC, Tim Berners-Lee, 2005

Semantic Web is an extension of the current Web in which information is given well-defined meaning …”
- Tim Berners-Lee, 2001
Is Linked Data Semantic Web in new clothes?

“Linked Data refers to a set of best practices for publishing and connecting structured data on the Web.”

“Semantic Web is an extension of the current Web in which information is given well-defined meaning…”

- Tim Berners-Lee, 2001

“Linked Data refers to a set of best practices for publishing and connecting structured data on the Web.”

- Chris Bizer, 2009

We need to “put the Web back in Semantic Web”

- ISWC, Tim Berners-Lee, 2005

“In the Semantic Web, it is not the Semantic which is new, it is the Web which is new”

- Chris Welty/IBM

11 Aug 2011

@jahendler
jahendler

NO NO NO NO NO NO NO NO - "one ontology to rule them all" is NOTHING to do with Semantic Web #greenapps

11 Aug via TweetDeck  ★ Unfavorite  → Retweet  Reply

Retweeted by ehCkzgUgJN and 13 others
Linked Data – so what?

“Life sciences will drive adoption of the Semantic Web, just as high-energy physics drove the early Web.”
- BioIT World, 2005
Integrate Data for better drugs

CHR

Patient2

Breast cancer

BRCA1

EGFR

Patient1

Entrez Gene

epidermal growth factor receptor

nih:has_description

nih:interacts

nih:inhibitedBy

nih:EGFR

nih:PCI

nih:EGF

Drugbank

Antineoplastic_Agents

db:category

db:target

db:chemicalStructure

db:EFGR

db:cetuximab

QVQLKQSGPGVLVPQSLSIT CTSVGFLTNYGHWVRQSP GKLWLGVWIWSGGNTDYLN PFTSRLINKDNSKQVFFKMN SLQSNQTYYCARALTYYD YEFAYWGQ
Integrate Data for better drugs

CHR
- Patient2
- BRCA1
- Breast cancer

Entrez Gene
- epidermal growth factor receptor
- nci:has_description
- nih:interacts

Drugbank
- Antineoplastic_Agents
- db:category
- db:target
- db:chemicalStructure
- db:cetuximab
- QVQLKQSGPGLVQPSQSLISIT
- CTVSFSNLLYNYGHVWQSP
- GKGLEWLGVIESGNLDSQVFFKM
- SLQSMNDTAIYYCARALTYYD
- YEAFAYWGQ

owl:sameAs
- nih:EGFR
- nih:PCI
- nih:EGF

nih:EGFR
- nih:inhibitedBy
- nih:interacts
Integrate Data for better drugs

CHR

Entrez Gene

epidermal growth factor receptor

nih:interacts

nih:has_description

nci:has_description
db:target

db:category

db:chemicalStructure

Drugbank

db:cetuximab

Breast cancer

Patient1

Patient2

BRCA1

db:EFGR

nih:inhibitedBy

nih:interacts

nih:PCI

nih:EGF

db:Antineoplastic_Agents

nih:inhibitedBy

帝王生长因子受体

nih:interacts

nih:has_description

nci:has_description
db:target

db:category

db:chemicalStructure

QVQLKQSGPGLVQPSQSLSIT
CTVSGFSLTYNYGHVRQSP
GKGLEWLVGIVSWGGNTDYNP
FTSRLSINSKNSKQVFDFKMN
SLQSNDTAIFYCARALTYDY
YEFAYWGQ
Linked Data – so what?

From Clinical Research

Safer and better drugs and treatments

Linked Data – so what?

Safer and better drugs and treatments

Cancer biology: Malaria drug shrinks tumours

*Nature 472*, 8 (07 April 2011) | doi:10.1038/472008b

Published online 06 April 2011

Subject terms: Cancer
The problem with drug discovery

- ~20,000 genes
- ~100 interesting genes/proteins
- ~10 interesting pathways
- ~5 proteins testable in the lab

High-throughput technologies

Linked Data

Computational statistics

Browse databases

Literature

Hypothesis Generation

“I like to call it low-input, high-throughput, no-output biology.”

Sidney Brenner
Linked Data for Understanding Cancer
Experimental results + Linked Data = Knowledge

Do these genes share a pathway?

What is the function of each of the cancer genes?

Is there a drug that targets all or most of those genes?
What diseases are genetically linked to Alzheimer’s?

Select samples from Alzheimer’s

Select relevant experiments with those samples

Select genes that are over-expressed

Use the genes to query the diseasome

PREFIX diseasome: <http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseasome/> ...

SELECT DISTINCT ?diseaseName ?geneLabel ?geneName WHERE {

  ?sampleList biordf:patients_have_disease ?alzheimers .
  FILTER (?alzheimers = doid:DOID_10652 )

  ?experimentSet dct:isPartOf ?microarray_experiment
  biordf:has_input_value ?sampleList ;
  biordf:differentially_expressed_gene ?gene ;
  biordf:has_ouput_value ?foldChange .


  FILTER (xsd:float(?foldChangeValue) > 0)

  SERVICE < http://www4.wiwiss.fu-berlin.de/diseasome> {
  }

}
### Some SPARQL federation results

<table>
<thead>
<tr>
<th>geneSymbol</th>
<th>associatedDisease</th>
<th>possibleDrug</th>
<th>trialDescription</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Cetuximab&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Gefitinib&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Lidocaine&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Sorafenib&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Erlotinib&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Letrozole&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Flavopiridol&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
<tr>
<td>&quot;EGFR&quot;</td>
<td>&quot;Adenocarcinoma&quot;</td>
<td>&quot;Trastuzumab&quot;</td>
<td>&quot;Condition #398 (Adenocarcinoma)&quot;</td>
</tr>
</tbody>
</table>
An app for Tracking superbugs
Linked Data Best Practices

1. Unstructured data (text documents)
2. Structured data (XML documents)
3. Manual input
4. Structured data (relational tables)
5. Ontology
6. SPARQL endpoint
7. Semantic web application
SPARQL as we know it

=> Not very exciting...
SPARQL as it SHOULD work

Gene

Diseasome

Drugbank

Clinical Trials
Problems in Web of Data land…

- SPARQL endpoint downtime

http://labs.mondeca.com/sparqlEndpointsStatus/
Problems in Web of Data land…

- URI that do not dereference

- Orphan URI

- No updates… no real incentive to keep the work going beyond POC
Most BioData is Still in Silos

- With Linked Data, we’ve reached only the tip of the iceberg
How a semantic web API for genomics became possible
SPARQL as SQL

- Converting the query instead of converting the data can create very ugly queries…

- We also want to be good citizens – automated RDB2RDF tools don’t give us control over the data so that we can specify ontologies

- It is also not straightforward to link to external URI, for example, bio2rdf
How about SPARQL?

- What if we could use SPARQL to
  - Align the schemas/ontologies
  - Create links between URI
  - Transform values as needed

- And we want to do all this without forcing the data into a triple store, where it would likely become outdated very quickly?
SWObjects Workshop!