The Integration of Biological Data Using Semantic Web Technologies

Susie Stephens
Principal Product Manager, Life Sciences
Oracle

susie.stephens@oracle.com
Outline

- Complexity of Biological Data
- Oracle’s RDF Data Model
- Life Sciences Use Cases
The Complexity of Biological Data
Pharmaceutical Productivity

$897 million

including post-approval R&D costs
to develop a new prescription drug
= 250% increase in a decade
  • Inflation-adjusted
  • Including failures

Source: PhRMA & FDA 2003
RDF Triples in Life Sciences

Subject
<Protein A>  <interacts_with>  <Protein B>

Predicate

Object

<Protein A>

<interacts_with>
<located>
<participates_in>
<influences>

<Cell A>
<Pathway A>
<Disease A>
The Semantic Web Vision

Source: Stephens et al. J Web Semantics 2006
Outline

- Life Sciences Data
- Oracle’s RDF Data Model
- Use Cases
Oracle and RDF: Motivation

- Customer requests
- RDF (and OWL) are maturing
- Oracle supports open standards
- Complements Oracle’s information management approaches
- Ability to leverage existing technologies
Oracle RDF Data Model

- Support for RDF and RDFS
- Object-relational implementation
- Subjects and objects are re-used
- Links represent complete RDF triples

RDF Triples:
- \{S_1, P_1, O_1\}
- \{S_1, P_2, O_2\}
- \{S_2, P_2, O_2\}
SPARQL-like Query Capability

- A table function allows a graph query to be embedded in a SQL query
- Searches for an arbitrary pattern against the RDF data
- Includes inferencing based on RDF, RDFS, and user-defined rules
Enterprise Functionality

- Real Application Clusters (RAC), Security
- Multi-threaded, parallel processing, indexed, etc.
- Performance testing with UniProt

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 M Triples</td>
<td>0.86</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>0.18</td>
<td>0.46</td>
</tr>
<tr>
<td>20 M Triples</td>
<td>0.95</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>0.19</td>
<td>0.47</td>
</tr>
<tr>
<td>40 M Triples</td>
<td>0.96</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>0.18</td>
<td>0.47</td>
</tr>
<tr>
<td>80 M Triples</td>
<td>1.03</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>0.20</td>
<td>0.49</td>
</tr>
<tr>
<td>Maximum σ</td>
<td>0.054</td>
<td>0.002</td>
<td>0.002</td>
<td>0.011</td>
<td>0.065</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Units in seconds

Source: Chong et al. VLDB 2005
Image Search

“Find me all DICOM images that contain the term ‘Jaw’”

- Map relationships to terms using RDF triples
  - ‘Mandible’, sameAs’, ‘Jaw’
  - ‘Maxilla’, ‘partOf’, ‘Jaw’
Text Search

“Find me all papers that contain the term ‘Jaw’”

• Map relationships to terms using RDF triples
  - ‘Mandible’, sameAs’, ‘Jaw’
  - ‘Maxilla’, ‘partOf’, ‘Jaw’
Data Integration

- **SQL / RDBMS**
  - Concise, efficient transactions
  - Transaction metadata is embedded or implicit in the application or database schema
- **XQuery / XML**
  - Transaction across organizational boundaries
  - XML wraps the metadata about the transaction around the data
- **SPARQL / RDF**
  - Information sharing with ultimate flexibility
  - Enables semantics as well as syntax to be embedded in documents
Download the Database!

Oracle Database Enterprise Edition 10g Release 2

Outline

• Life Sciences Data
• Oracle’s RDF Data Model
• Use Cases
Currently, PKB contains data from the following databases with the following species:

- **Human** (979 pathways)
  - Reactome (705 pathways)
  - BioCyc (166 pathways)
  - KEGG (108 pathways)
- **E. coli** (489 pathways)
  - Reactome (75 pathways)
  - BioCyc (191 pathways)
  - KEGG (223 pathways)
- **Yeast** (599 pathways)
  - Reactome (62 pathways)
  - BioCyc (364 pathways)
  - KEGG (173 pathways)

Source: http://pkb.stanford.edu/
Eli Lilly Use Case

Discovery TAT Service-Oriented Architecture

“Pluggable Architecture”

Discovery TAT

Semantic Integration

MESH

Gene Ontology

Reference IDs

Pathways

Visual Studio 2005/NET 2.0

Microsoft Patterns & Practices Block

Composite Application Block

CAB Module Handling/Hosting

CAB Event/State Handling

Common XML Schemas (XSD)

WSDL Web Service

System A

Biology

System B

Chemistry

System C

Toxicology

... Database N

University of Texas Health Science Center Use Case

- Time
- Upper Ontology SUMO
- Event
- Geospace
- Disease/Syndrome
- BioAgent
- UMLS
- Signs/Symptoms
- PHIN
- Bioterrorism
- WMD

Image Source: Semantic Technologies Conference 2006
BioRDF

**Active Tasks**
- /Reagents (Status)
- /SenseLab
- /Using SW Technologies to Find Small Molecules that Bind to Proteins
- /Gene Neural related gene data
- /OMIM Neural diseases
- /Natural Language Processing and RDF
- /Ligand-Receptor Interaction, Molecular Interaction Networks, Ontology Evolution
- /Vocabulary Requirements

**Proposed Tasks**
- /Brain Connectivity
- /Brain Atlas condition to scans
- /Protein Neural related protein data
- /Ruby On Rails and ActiveRDF

Source: http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup
Summary

- The Semantic Web provides the ability to more easily integrate heterogeneous data
- Oracle has a scalable, secure, highly-available RDF Data Model
- Adoption of Semantic Web technologies is accelerating
- Make your data sharable, make it available in RDF