Open Terminology Portal (TOP)

Frank Hartel, Ph.D.
Associate Director, Enterprise Vocabulary Services
National Cancer Institute, Center for Biomedical Informatics

NCBO Scientific Council
October 16, 2007
NCI BioPortal and the TOP Project Links

- NCI BioPortal Project
  https://gforge.nci.nih.gov/projects/lex-browser/
- Open Terminology Portal Project
  https://gforge.nci.nih.gov/projects/openportal/
Why did NCI find BioPortal so interesting?

- Our caCORE and caGrid infrastructure relies on terminology servers to provide access to the base semantics used to construct metadata and data model semantics.
- During 2007 we have moved away from our legacy terminology server architecture to an open architecture built around LexGrid (LexBIG).
- Our legacy terminology browsers are not compatible with the LexBIG infrastructure – but BioPortal largely was and we needed a interactive Web tool for our users.
- BioPortal had a simple, effective GUI.
- While NCBO had not built it for distribution, they were willing to let us have the code.
Why did NCI not simply use NCBO BioPortal?

- We needed to serve a Metathesaurus (NCI Metathesaurus – similar to UMLS Metathesaurus), not only individual ontologies
- Our operations model precludes uploading of ontologies by end users
- Some of the terminologies we serve have license restrictions prohibiting redistribution or requiring license validation before downloading
- Our infrastructure is not the same as the one that NCBO BioPortal assumed
  - Database is MySQL
  - Index files, database and LexBIG not co-located with BioPortal
  - NCI BioPortal bound to caCORE/LexBIG product and its release cycle
    - Planning, engineering methodology and artifacts, documentation QA, training
    - We use Distributed LexBIG API, not local LexBIO API
Context of Open Development – NCI Enterprise Vocabulary Services

EVS Product

<table>
<thead>
<tr>
<th>Legacy Operations</th>
<th>TDE</th>
<th>NCIt Releases</th>
<th>DTS</th>
<th>DTS-RPC</th>
<th>caCORE 3.2</th>
<th>NCI Term Browser</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEME</td>
<td></td>
<td>NCI Meta Releases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Emerging Infrastructure Operations

<table>
<thead>
<tr>
<th>Open Content Development</th>
<th>NCIt Releases</th>
<th>Other Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiomedGT</td>
<td>NCI Meta Releases</td>
<td>UMLS Meta Releases</td>
</tr>
</tbody>
</table>

Other open ontologies

Infrastructure Development

<table>
<thead>
<tr>
<th>Semantic Media Wiki</th>
<th>NCI Protégé/OWL</th>
<th>Classification Services</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Workflow</td>
</tr>
</tbody>
</table>
What did we do with BioPortal to date?

- We tried to make BioPortal less NCBO specific
  - Added MySQLSequenceHandler and other code to support MySQL
  - Removed Oracle specific dependencies
  - Removed NCBO database and relevant code segments
  - Removed Apache HTTP server & Tomcat Connector
  - Integrated JBoss with AT&T Graphviz
    - (All Graphviz output files including .dot, .jpg, and .map) are now generated in a JBoss tmp subdirectory instead.
- Modified Windows-specific code; deployable on Linux
- Integrated with distributed LexBIG
What did we do with BioPortal to date? - ctd

- Metadata files loaded to MySQL via LexBIG metadata loader
- Removed code specific NCBO workflow (user registration, ontology submission, admin functions)
- Created Quick and Advanced Search Tabs
  - Search by source, class property, association, etc.
- Removed EJB3 related to NCBO database
- GUI changes -- Examples
  - Library statistics read from MYSQL
  - Metathesaurus browse and display
  - On line help, user guide
## NCI BioPortal Library Statistics

### Ontologies

<table>
<thead>
<tr>
<th>Name</th>
<th>Format</th>
<th>Current Version</th>
<th>Content Location</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>test2</td>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneOntology.owl</td>
<td>??</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL7.owl</td>
<td>v3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>autos</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The MGED Ontology</td>
<td>classified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Drug File - Reference Terminology, NCI_072605</td>
<td>NCI_072605</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCI MetaThesaurus</td>
<td>RRF</td>
<td>200706</td>
<td><a href="http://ncicb.nci.nih.gov/core/EVS">http://ncicb.nci.nih.gov/core/EVS</a></td>
<td></td>
</tr>
<tr>
<td>UMLS Semantic Network</td>
<td>3.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNOMED Clinical Terms, 2006_07_31</td>
<td>2006_07_31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCI Thesaurus</td>
<td>OWL-FULL</td>
<td>07.08d</td>
<td><a href="http://ncicb.nci.nih.gov/core/EVS">http://ncicb.nci.nih.gov/core/EVS</a></td>
<td></td>
</tr>
<tr>
<td>GMP</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Browse in individual ontologies as on NCBO BioPortal
Browse in Meta requires selection of a source
In Meta, graph views also require source selection.
Genesis of TOP

- **NCI changes to BioPortal**
  - Made BioPortal code easier to host locally, but…
  - Forked code – need to reintegrate!
  - Demonstrated interest in local instances of BioPortal to meet local needs
- **Others with interest in BioPortal emerged** – Mayo Clinic and UK Cancer Grid beside NCI and NCBO
- **Formed initial partners in Open Terminology Portal Project**
  - Open development, open source, participant driven governance still under development
  - Additional partners are desired
Current TOP Activities

- **Early days…**
- **Charter adopted**
- **Discussing open development process and project management**
- **Scope definition for near term “de-forking”**
  - Initial goal is to continue to refine the BioPortal code and close the fork
    - Requirements and design goals, architecture under development
    - Close the fork, robust, enterprise scale, site neutral, etc.
- **Requirements gathering for future TOP offerings, initiatives**
  - A bit further in the future, development of services, components and other resources to address patterns of terminology use are being discussed
  - **Target architecture**
    - Architectural considerations discussions beginning