Basic Premise

• It is both *useful* and *practical* to expose caBIG data sets as Linked Data.
What is Linked Data?

- **Linked Data**
  - Set of principles/best practices for publishing data on the Web.
  - Aligned with WWW architecture
  - Web of Data vs. Web of Documents

- **Semantic Web**
  - Web of machine interpretable data
caBIG+Linked Data is Useful

- Immediately accessible through Web
- Interlinking with other data sets creates additional value
- Enables powerful technologies
  - linked data browsers
  - semantic search engines
  - logic-based reasoners
- Discovery of new patterns in data
caBIG+Linked Data is Practical

• Aligns with caBIG goals
  • Open, federated data sharing network
  • Precise semantic definitions to enable interoperability

• Semantic Infrastructure Reuse
  • Reuse existing processes and tools
  • Minimize barriers (cost) to data providers

• Linked Data is gaining momentum
  • Network effect -> increased value, stable technology
Outline

• Background
• Linked Data & caBIG
• Preliminary Work
• Discussion
Linked Data Background

- Title of a design note by Tim Berners-Lee
- Best practices for publishing data on the Web.
- Web of data using HTTP, URIs, and typed (RDF) links (vs. hypertext links).
Linked Data Principles

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up for those names
3. When someone looks up a URI, provide useful RDF information
4. Include RDF statements that link to other URIs so that they can discover related things
Example: Navigating caBIG Web
caArray -> caTissue -> COPPA

https://array.nci.nih.gov/caarray/project/abate-00279
Example: Navigating caBIG Web
caArray -> caTissue -> COPPA

https://array.nci.nih.gov/caarray/project/abate-00279

GET /caarray/project/abate-00270 HTTP/1.0
Accept: application/rdf+xml
Example: Navigating caBIG Web
caArray -> caTissue -> COPPA

https://array.nci.nih.gov/caarray/project/abate-00279

GET /caarray/project/abate-00270 HTTP/1.0
Accept: application/rdf+xml

caarray1:abate-00279

caarray1:Experiment

"Activator protein-1 transcription factors are ..."

catissueX:specimen123

catissueY:specimen456
Example: Navigating caBIG Web
caArray -> caTissue -> COPPA

https://catissue.some.org/resource/specimen123

diagram:
- caarray1:abate-00279
  - rdf:type -> caarray1:Experiment
  - dc:title -> "Activator protein-1 transcription factors are ..."
- caarray:source
  - caarray:source
    - caarray:source
      - catissueX:specimen123
    - catissueY:specimen456
Example: Navigating caBIG Web
caArray -> caTissue -> COPPA

https://catissue.some.org/resource/specimen123

GET /resource/specimen123 HTTP/1.0
Accept: application/rdf+xml

caarray1:abate-00279

caarray1:source

"Activator protein-1 transcription factors are ..."

caarray1:source

catissueX:specimen123

catissueY:specimen456
Example: Navigating caBIG Web
caArray -> caTissue -> COPPA

https://catissue.some.org/resource/specimen123

GET /resource/specimen123 HTTP/1.0
Accept: application/rdf+xml

"Activator protein-1 transcription factors are ..."
Example: Navigating caBIG Web caArray -> caTissue -> COPPA

"Activator protein-1 transcription factors are ..."
More about Linked Data

- **De-referencing URIs**
  - Non-information vs. information resources
  - Hash URIs vs. slashes
  - 303 redirects & content negotiation
- **Vocabulary of Interlinked Datasets (voiD)**
  - What is the subject of the dataset?
  - Where is the SPARQL endpoint?
  - Relationships with other datasets.
Who is providing Linked Data now?

- Publish existing open license datasets as Linked Data on the Web
- Interlink “things” between different data sources
- Size of dataset: approx. 7.8 billion triples
- Number of links: approx. 143 million
Life Sciences Contributors

- HCLS Linked Open Drug Data
  - Won 2009 Triplification Challenge
- Bio2RDF
  - 40 biology-, gene- and medical-related datasets (altogether 2.3 billion triples)
- Many more…
  - http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSetS
Many are getting involved…

CNET Partners with Thomson Reuters on Linked Data Initiative

The latest implementation of OpenCalais, the Semantic API by media company Thomson Reuters, has just been announced. It’s with new media stalwart CNET, which has signed up to use OpenCalais for semantic analysis of its tech product reviews, news, and blog posts. CNET has also joined Thomson Reuters as one of the first commercial media companies to publish its data to the Linked Data community on:
caBIG+Linked Data: How?

- Need RDF vocabularies for describing caBIG data sets.
- Need consistent approach to naming things with URIs.
- Need to *RDFize* caBIG datasets.
  - Must minimize technology barriers to data providers.
RDF Vocabularies

- UML-OWL Generator [1]
- OWL representations of information models
- Retains semantics of original model
- Includes NCIt concepts
- Provides schemata for data

RDF Vocabularies

Diagram:
- Domain Model Ontology
- Semantic Metadata Ontology
- NCI Thesaurus (NCIt)
- caGrid Metadata Ontology
- NCIt Module for Data Service Ontology
- Data Service Ontology
Naming caBIG resources with URIs

• caGrid Identifier Framework
  • Addresses issues of change.
  • Based on Persistent URLs (PURLs)
  • Provides HTTP URI Naming Authority, Prefix Authority, and resolution scheme.
  • Consistent with Linked Data principles.
Naming caBIG resources with URIs

1. data object Y
2. data object metadata
3. identifier X
4. identifier X
5. redirect
6. data object metadata
7. retrieve
8. data object Y

Identifier Curator

Naming Authority (foo.osumc.edu)
prefix = http://na.cagrid.org/foo

Prefix Authority (na.cagrid.org)
/foo => http://foo.osumc.edu

Data Owner

Identifier X = http://na.cagrid.org/foo/8576
RDFizing caBIG Data

- Alternative approaches
  - Static RDF files
  - Native triplestore
  - Adapter over API (e.g. Data Service API)
  - Adapter over relational DB
Recommended Approach

• Adapter over relational DB
• Factors:
  • caCORE SDK generated services use relational database backend.
  • Potentially large data sets.
  • Potential high frequency of change.
Initial Process

1. UML-OWL Generator -> OWL info. model
2. D2R Service -> Generate RDF-relational mapping
3. Use caCORE SDK XMI to modify D2R mapping to use OWL classes representing information model.
4. D2R Server to exposing Linked Data & SPARQL HTTP Interfaces
Preliminary Work

• Exposing caTissueSuite 1.1 data set as Linked Data.
• Goal: To validate high-level use cases with important, real caBIG data set.
caTissue Suite 1.1 OWL
caTissue Suite 1.1 OWL

The image shows a screenshot of the caTissue Suite 1.1 OWL interface, which is a tool for working with the Open Web Ontology Language (OWL) in the context of biomedical informatics. The interface is used to define and visualize ontologies, which are formal representations of a set of concepts within a domain and the relationships between those concepts.

The specific graph in the screenshot highlights a class `SpecimenEventParameters` and its properties. The properties include attributes such as `ThawEventParameters_comment`, `ThawEventParameters_id`, and `ThawEventParameters_timestamp`. The diagram also shows inferred anonymous superclasses and subproperties related to the `SpecimenEventParameters` class.
D2R Mapping

- Maps relational model to RDF classes
- Mapping expressed in RDF (N3 syntax)
- ClassMap
  - Maps tables to RDFS or OWL classes
- PropertyBridge
  - Maps columns to simple RDF properties (i.e. OWL DatatypeProperty)
  - Maps FKs to RDF links (OWL ObjectProperty)
- Both tables and columns can be mapped to existing RDFS or OWL vocabularies.
@prefix vocab: <http://localhost:8080/owl/catissue1_list.owl/> .
...
map:catissue_tissue_specimen a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern ...
  d2rq:class vocab:TissueSpecimen;
  ...

map:catissue_specimen_event_param_SPECIMEN_ID a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:catissue_specimen_event_param;
  d2rq:property vocab:SpecimenEventParameters_specimen_AbstractSpecimen;
  d2rq:refersToClassMap map:catissue_abstract_specimen;
  d2rq:join "catissue_specimen_event_param_SPECIMEN_ID =>
    catissue_abstract_specimen.IDENTIFIER";
  ...

Linked Data & SPARQL Interface

**SPARQL:**
```sparql
PREFIX db: <http://localhost:2020/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX map: <file:/Users/joshua/packages/d2r-server/d2r-server-0.7/catissue2_mapping.n3#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vocab: <http://localhost:8080/owl/catissue1_1list.owl/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?instance
WHERE { ?instance a <http://localhost:8080/owl/catissue1_1list.owl/TissueSpecimen> }
ORDER BY ?instance
```

Results: Browse Go! Reset

All instances of class http://localhost:8080/owl/catissue1_1list.owl/TissueSpecimen:

<table>
<thead>
<tr>
<th>Instance</th>
</tr>
</thead>
<tbody>
<tr>
<td>db:catissue_tissue_specimen/1007</td>
</tr>
<tr>
<td>db:catissue_tissue_specimen/1008</td>
</tr>
<tr>
<td>db:catissue_tissue_specimen/1009</td>
</tr>
<tr>
<td>db:catissue_tissue_specimen/1010</td>
</tr>
<tr>
<td>db:catissue_tissue_specimen/1011</td>
</tr>
</tbody>
</table>

**Linked Data navigation**

**OWL Class defined in Information Model**

**SPARQL Query Interface**
### Linked Data & SPARQL Interface

**Description of http://localhost:2020/resource/catissue_abstract_specimen/1007:**

<table>
<thead>
<tr>
<th>property</th>
<th>hasValue</th>
<th>isValueOf</th>
</tr>
</thead>
<tbody>
<tr>
<td>vocab:catissue_abstract_specimen_IDENTIFIER</td>
<td>1007</td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_abstract_specimen_INITIAL_QUANTITY</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_abstract_specimen_LINEAGE</td>
<td>&quot;New&quot;</td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_abstract_specimen_PATHOLOGICAL_STATUS</td>
<td>&quot;Malignant&quot;</td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_abstract_specimen_SPECIMEN_CHARACTERISTICS_ID</td>
<td>db:catissue_specimen_char/898</td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_abstract_specimen_SPECIMEN_CLASS</td>
<td>&quot;Tissue&quot;</td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_abstract_specimen_SPECIMEN_TYPE</td>
<td>&quot;Fixed Tissue&quot;</td>
<td></td>
</tr>
<tr>
<td>rdf:type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rdfs:label</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_abstract_specimen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vocab:catissue_specimen #1007</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Mapped role defined in OWL Information Model*
Prototype Evaluation

- Very preliminary work
- Relatively straightforward to map tables and columns to OWL classes and properties.
- Similar functionality to caCORE GetXML or GetHTML HTTP APIs.

Challenges

- Interpreting alternative O-R mapping strategies for OO Inheritance.
- D2R server Performance
Related Work in caBIG

- Prostate Cancer Information System (PCIS)
- A prototype system (Fox Chase Cancer Center)
  - Developed Prostate Cancer Ontology (PCO)
  - Apply PCO to integrate two database systems
    - Tumor Registry
    - Prostate Cancer Database
  - A web-based ontology query formulation

- Hua Min, Frank J. Manion, Elizabeth Goralczyk, Yu-Ning Wong, Eric Ross, J. Robert Beck *Integration of Prostate Cancer Clinical Data Using an Ontology* (JBI 2009)
PCIS Ontology Mapping

A. Prostate Cancer Ontology
   - Person
     - Name
     - ID
     - Sex
     - Age
     - Address
     - MRN

   - Patient
   - Biox
   - Diagnosis

B. Prostate Cancer Database
   - V_DEMOP_PRCA_RADONC
     - ID
     - MR
     - Name
     - Sex
     - Age
     - Address
   - BX
     - ID
     - dt

C. Tumor Registry
   - ORA_PT
     - PT_Accn_No
     - PT_Med_Rec_No
   - ORA_DG
     - DG_Accn_No
     - DG_Seq_No
     - DG_Site

Mapping
Other RDFizer Tools

- W3C RDB2RDF Incubator Group
  - Surveyed all existing approaches
  - Developed comparison framework
  - Recommends using Rule Interchange Framework (RIF)
  - No existing tools support RIF

http://www.w3.org/2005/Incubator/rdb2rdf/
Performance Analyses

• Benchmarking Tests Exist
  • Lehigh University Benchmark (LUBM)
  • Ontology Benchmark (UOBM)
  • Berlin SPARQL Benchmark (BSBM)

http://esw.w3.org/topic/RdfStoreBenchmarking
Berlin SPARQL Benchmark

- Relational still faster than RDF & SPARQL

<table>
<thead>
<tr>
<th>Dataset Size 25M</th>
<th>Number of clients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Sesame</td>
<td>1,343</td>
</tr>
<tr>
<td>Jena TDB</td>
<td>353</td>
</tr>
<tr>
<td>Jena SDB</td>
<td>968</td>
</tr>
<tr>
<td>Virtuoso TS</td>
<td>4,123</td>
</tr>
<tr>
<td>Virtuoso RV</td>
<td>12,972</td>
</tr>
<tr>
<td>D2R Server</td>
<td>140</td>
</tr>
<tr>
<td>MySQL</td>
<td>18,578</td>
</tr>
<tr>
<td>Virtuoso SQL</td>
<td>69,585</td>
</tr>
</tbody>
</table>

http://www4.wiwiss.fu-berlin.de/bizer/BerlinSPARQLBenchmark/results/index.html
Discussion

• Limitations of mappings
• Link discovery and maintenance
• Performance
  • Enhancements/optimizations are needed
  • RDB2RDF tools show promise
    • Need to exploit research experience in RDBMSs, esp. in query translations
References

• A Survey of Current Approaches for Mapping of Relational Databases to RDF – W3C RDB2RDF Incubator Group Report http://www.w3.org/2005/Incubator/rdb2rdf/
• Bizer et al. Linked Data – the story so far.