Publishing Linked Open Data: Experiences with ESSI-LOD Project

Eric Rozell and Tom Narock
Outline

• RDF Vocabularies
• Mobile App Software Stack
• Linking Data
RDF Vocabularies Used
RDF Vocabularies Used
RDF Vocabularies Used

- TWC - http://tw.rpi.edu/schema/
- SWRC - http://ontoware.org/swrc/
- GEO - http://www.w3.org/2003/01/geo/
- SKOS - http://www.w3.org/2004/02/skos/
Mobile App Software Stack

• jQuery Mobile
• LODSPeaKr
• OpenLink Virtuoso
• ESSI-LOD Pipeline
• AGU “Meeting-at-a-Glance”
## AGU “Meeting-at-a-Glance”

### Session Information

#### 2011 Fall Meeting

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
<th>Title</th>
</tr>
</thead>
</table>
| 0800  | IN11A   | Halls A-C (Moscone South) | Creating Decision Support Products in a Rapidly Changing Environment Posters (joint with C. V)  
Presiding: K Keiser, University of Alabama in Huntsville; S Ebersole; D M Hardin, Univ of Ala Huntsville  
Print-friendly Session Details |
| 0800  | IN11B   | Halls A-C (Moscone South) | Current Capabilities and Future Needs of Near Real Time Data I Posters (joint with A, B, C, NH, OS, V)  
Print-friendly Session Details |
| 0800  | IN11C   | Halls A-C (Moscone South) | Cyberinfrastructure That Advances Understanding of Ecosystem Processes Posters (joint with B, EP, GC)  
Presiding: C E Tweedie, University of Texas at El Paso; J A Gamon, Univ of Alberta  
Print-friendly Session Details |
| 0800  | IN11D   | Room 102 (Moscone South)  | High-Resolution Modeling in the Geosciences Using GPU and Many-Core Architectures I (joint with A, GP, GC, NS, NG, OS, S, Di, T, V)  
Presiding: D A Yuen, Univ of Minnesota; A Schultz, Oregon State Univ; M G Knepley, University of Chicago  
Print-friendly Session Details |

### Monday Morning 2

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
<th>Title</th>
</tr>
</thead>
</table>
| 1020  | IN12A   | Room 102 (Moscone South)  | High-Resolution Modeling in the Geosciences Using GPU and Many-Core Architectures II  
Presiding: D A Yuen, Univ of Minnesota; A Schultz, Oregon State Univ; M G Knepley, University of Chicago  
Print-friendly Session Details |

### Monday Afternoon 1

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
<th>Title</th>
</tr>
</thead>
</table>
| 1340  | IN13A   | Halls A-C (Moscone) | High-Resolution Modeling in the Geosciences Using GPU and Many-Core Architectures III Posters (joint with A, GP, GC, NS, NG, OS, S, Di, T, V)  
Print-friendly Session Details |

*XML data coming (from AGU) soon!!!*
ESSI-LOD Pipeline

Source code available from: http://essi-lod.org
OpenLink Virtuoso

• Pros
  – All-in-one SPARQL endpoint and triple store
  – Open source: http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/
  – Scales better than TDB+Joseki
  – Good support for named graphs
  – Can be optimized for free-text search

• Cons
  – Free-text search only scales so far...
  – More difficult to configure than other triple stores
  – Interesting hard-coded limit on group graph patterns (57... why not 42?)
  – Requires non-standard SPARQL constructs for simple reasoning tasks, e.g.:
    • ... ?s skos:broadMatch ?o OPTION(transitive) ...
    • ... DEFINE input:inference “myInferenceGraph” ...
• Awesome tool for quickly publishing linked data from SPARQL endpoint
• Uses model-view paradigm
• Uses Haanga PHP templates to create virtually any type of REST service based on SPARQL responses (e.g., JSON, RDF/XML, turtle, HTML+RDFa)
• http://lodspeakr.org
• LODSPeaKr TWED talk:
  – http://www.ustream.tv/recorded/20630480
• Targeted towards development of mobile web apps
• Could use tools to convert to native app (... as with any website)
• Provides APIs for standard mobile events covering multiple mobile devices
• http://jquerymobile.com/
• For ESSI-LOD, we embedded jQuery mobile scripts directly in our LODSPeaKr views
Linking Data

• DBPedia Spotlight
• Crowd-sourcing
• Google Refine
• Google Geocoding API
• Clique Analysis
• Annotates abstracts with DBPedia terms
• Pros
  – Unlimited free use (open source and open data)
  – Links plain text content to DBPedia URIs
  – Can use web service or host locally
  – Can arbitrarily constrain types of annotations using SPARQL
• Cons
  – Web service is slow (takes a few days to run 100,000s of abstracts)
  – False positive rate is very high
• [http://spotlight.dbpedia.org](http://spotlight.dbpedia.org)
Crowd-Sourcing

• Disambiguate authors and affiliations

• Pros
  – Structured data disambiguation
  – Minimize false positives

• Cons
  – Need to “market” the service
  – Even well marketed, does not scale
  – What about trust/authentication?
Google Refine

- [http://code.google.com/p/google-refine/](http://code.google.com/p/google-refine/)
- Disambiguate authors and affiliations

**Pros**
- Simple to clean up large amounts of data
- Clustering features “for free”

**Cons**
- Semi-automated
- Spreadsheet-based, cannot account for connections across rows (e.g., social network)
Google Geocoding API

• Disambiguate affiliations from lat/long
• Pros
  – Fully automated
• Cons
  – Does not handle sub-organizations
  – Limits the number of calls per day
  – Web services are slow for big data
Clique Analysis

• If two people/organizations with a very similar name belong to the same clique, they are probably the same entity

• Pros
  – Fully-automated...
    • within a degree of certainty

• Cons
  – Non-trivial implementation
Planning Ahead

• Google Refine + Geocoding = ?
• Specifically targeted Spotlight annotations
  – Locations
  – Organizations
• Web services for author and affiliation disambiguation
  – Like sameas.org for ESSI community
  – What are some common data formats for person/affiliation data? Mailing lists? CSVs?
Questions?