Resource Discovery for Extreme Scale Collaboration (RDESC)

- Karen Schuchardt (PNNL)
- Peter Fox (RPI)
- Alan Chappell, Sinan al-Saffar, Sumit Purohit (PNNL)
- Eric Rozell, Benno Lee (RPI)
Goal: Enable Broader Science Collaboration through Improved Resource / Knowledge Discovery

Research and develop scalable, open mechanisms for publishing, annotating, and linking resources to demonstrate scalable, hybrid **semantic** discovery

**Requirements**
- Provider-independent discovery at the level of science questions
- Small and large scale publishing
- Local control over data
- Flexibility in dialects
- Annotation and cross linking
- Scalable stores
Some current work

► **Policy:**
  - America Competes Act, NSF Policy on Research Products, FASTR bill…

► **General Data Publishing/Aggregation**
  - Data Citation Index
  - Linked Open Data
  - DataCite – find and reuse scientific “information”
  - DataUP – document, share tabular data
  - World Data System
  - UC’s eScholarship (Digital libraries)

► **Targeted Publishing/Aggregation**
  - Data Observation Network for Earth (DataOne)
  - Pangea (pangea.de) – Publisher: Earth and Environment science data
  - Earth System Grid (ESG)
  - Global Change Master Directory (GCMD)
  - Many….

► **Missing:** open, semantics, automatable, publishing tools

---

Data Citation Index: 500 repositories estimated to hold millions of GB of data. Although organic growth of repositories allows for great innovations, the variance also lead to a "silo" environment of disconnected resources.

LOD (2012: 31 billion triples, 500m links)
Methodology

Open-World: Evolve, iterate, design, deploy

Use Case

Analysis

Data Collection, Publication, Annotation, Linking Tools/Methods

Develop/Apply Ontology Models

Query Optimization

Review/adopt technology approach

Science/expert review & iteration

Prototype Discovery Interfaces
Use Case - Climate

Goal: Find observational data that can be used to validate new methods I’m developing in a land surface model.

► Find data sets that contain surface *essential climate* at daily temporal resolution from 1990-2000.

► Find observational data for soil moisture at depths to 30cm and that has corresponding meteorological data and information on the vegetation near the instruments
  ■ What protocol was used to collect and process the data
  ■ What organization(s) are involved
  ■ What else has the data been used for (e.g., publications)

► Find soil moisture data sets that have annual data between 1980 and 2000 that can be gridded to 20 km resolution
Use Cases (Continued)

Goal: Find detectors designs and performance information that I can apply to my new instrument

- What *portable* particle detectors or combination of detectors can detect gammas in the 50-500 keV range, and sensitive to both gammas and protons.

Goal: Find data processing algorithms and tools that have been applied to instrument X’s data that I could apply to my instrument data

Goal: Show what data, methods and tools were used to generate this image.
Climate Data Collection

- OPeNDAP
- NASA Echo system
- Global Change Master Directory
- Mercury consortium
- ARM
- Evaluating slew of other “services”

Triples: 400 M  
Data Files: 1.1M  
Data Sets: ~10k  
Sites: 24
Ontologies – a fractal view

Alignment and Mapping

Earth, Science, HEP, Domain

RDFS, Dublin Core, FOAF, SKOS, SPARQL

Time, OWL

Climate Science, SWEET

EarthCube, HOD, ESG, Essential Climate, CF, SWEET
Ontologies

Ontologies in current testbed

- Dublin Core
- GeoSPARQL
- ESV, VTSO
- ECHO
- Owl Time
- CF to RDF mapping
- RDF
- DataSet (5 varieties)

- Other (not used yet)
- SWEET
- HYDRO extensions to SWEET

SKOS Assertions

- skos:exactMatch
- skos:closeMatch
- skos:broadMatch
- skos:narrowMatch
- skos:relatedMatch

Essential Climate Variables, CF, VSTO
Data Store Scaling

- Triple stores have known issues with scaling:
  - Open source: scale to $10^8$
  - Commercial: scale to $10^9$
  - Simple, limited crawl: $10^{10}$

- Approach: exploit triple store where that technology is most effective:
  - Focus on typed paths and patterns.
  - Identify typical queries from use cases.

- Hybrid technology approaches to enhance scaling:
  - Reduce size of info in semantic store:
    - Order of magnitude (reification)
    - More from selective use of properties
  - Indices tailored to key search criteria
  - Evaluate map-reduce possibilities against use case queries (graph partitioning)
## Open Source Data Store Technology

<table>
<thead>
<tr>
<th>Feature</th>
<th>Virtuoso OpenSource (6.x)</th>
<th>Virtuoso Comm. (6.x)</th>
<th>USekM</th>
<th>Parilament</th>
<th>Strabon</th>
<th>BigData</th>
<th>4Store</th>
<th>ontotext</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGC Compliant</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>(Yes: With UseekM)</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Support for geoSPARQL1.0</td>
<td>NA</td>
<td>NA</td>
<td>Almost fully supported (as per project forum)</td>
<td>Yes</td>
<td>Subset of geoSPARQL</td>
<td>(Yes: With UseekM)</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Geometry Object Support</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>(Yes: With UseekM)</td>
<td>find Point in 2D</td>
<td></td>
</tr>
<tr>
<td>Coverage of Geometric Object Support</td>
<td>NA</td>
<td>Only Point</td>
<td>Polygon</td>
<td>Polygon</td>
<td>Polygon</td>
<td>Polygon</td>
<td>lat-lon based</td>
<td></td>
</tr>
<tr>
<td>Support for WKT</td>
<td>NA</td>
<td>NA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>(Yes: With UseekM)</td>
<td>No(proprietary/luce based free text search)</td>
<td></td>
</tr>
<tr>
<td>Temporal Index Support</td>
<td>NA</td>
<td>NA</td>
<td>Yes</td>
<td>Work in Progress</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query Processor Support</td>
<td>Sesame or Jena</td>
<td>Sesame or Jena</td>
<td>Sesame or Jena</td>
<td>Sesame</td>
<td>Yes</td>
<td>Sesame or Jena</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Last Release</td>
<td>2012-08-01</td>
<td>2012-08-01</td>
<td>1.2.0-a5(2013/2/4)</td>
<td>2.7.4(2012-11-09 )</td>
<td>3.2.3(2012/9/5/)</td>
<td>1.2.2 (2012/09/16)</td>
<td></td>
<td>OWLIM-SE is optimized for scalability</td>
</tr>
<tr>
<td>Scalability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Petabyte Scale</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Vision: Moving from isolated data silos and hidden data to open knowledge systems through open linking

Challenge – chicken and egg problem for data collection, metadata, annotation, linking, ontologies & dialects

Next Steps
- Data Collection
  - Mercury*, ARM, …
  - Ontology assignment
  - Annotation
- In depth evaluation of existing data stores
- Evaluate alternative ontology structures
- Complex Queries
- Initial hybrid implementation
- Initial search interfaces

Progress
- Initial Use Cases
- Analysis of data options
- Initial data collection
- Initial ontology mapping
- Surveyed existing semantic stores
- Installed 3 stores; 1 pending
- Ingested/initial query
- Participating in ESIP (Earth Sciences Information Partners)
Objective
Research and develop open mechanisms for describing, linking, searching, and **discovering** resources critical to large scale, multi-disciplinary science.

Approach
- Develop iterative, focused use cases
- Develop/employ/recommend publication mechanisms for use by both programs and people
- Implement a hybrid semantic store scalable to billions of pieces of knowledge
- Develop exploratory discovery user interfaces

Progress
- Defined use case(s):
  - climate modeling data discovery
  - HEP detector discovery
- Surveyed existing climate data providers and types of services
- Harvested 400M triples from ~25 sites
- Surveyed scalable triple stores with geospatial query capability; setting up test bed with top candidate technologies
- Defined Essential Climate Variables ontology
- Selected initial set of ontologies

Impact
- Increase opportunity for collaboration and ad hoc formation of new research teams by reducing reliance on personal contacts or manual information integration
- Reduce cost and time to deliver discovery tools tailored to different scientific communities