Semantic Curation Use Case

Jim Myers
myersj4@rpi.edu
Traditional Curation

- Very manual – researchers hand the problem to curators
- Only for the highest value data
- So much work to find data and metadata that assessment of value is ‘intuitive’

- SEAD seeks to ‘automate’ discovery and access
  - With that, we must now scale curation decision making
Sustainable Environment - Actionable Data

- An international resource for sustainability science
- Novel technical and business approaches to supporting the long-tail of research data
- Lifecycle support: actionable data services integrated with curation and preservation infrastructure

- $8M / 5 years + ?
  - First two years are prototype/pilot deployment
Sustainability Research is central to solving many of society’s most critical challenges

- Local processes aggregating to produce global consequences
- Multiple time scales
- Coupling of natural and human systems
- Interacting systems-of-systems requiring multidisciplinary understanding
  - Environmental – Economic - Social
Data Challenges in Sustainability Research

- Many dimensions, many coordinate systems, many scales, many formats, a long-tail of providers and users, ...
Long-tail Dynamics

- Sustainability research involves many small groups and limited central coordination
  - Storage, QA/QC, derived products, provenance, visualization are all hard problems for small groups
- Solutions must have low barriers, opt-in policies, and quick payback
Vision for the SEAD Active Content Repository and Services

- Deployment of capabilities to generate and capture metadata during active research with emphasis on providing value-added services as incentives will drive good curation practices

- Ingest costs for long-term preservation can be lowered through curator services leveraging active context/content

- Integration services enabled by a shared metadata/data repository will accelerate and transform sustainability research
SEAD CI

Semantic Content Middleware over Scalable File System and Triple Store

Flickr-style web management of data

The VIVO ‘Facebook for Science’ research information repository

Geospatial, social network mash-ups, workflows and services

Sensor data

Curation Services to harvest and package specific data sets

Federation of OAI repositories for long-term preservation
What makes data ‘curation-worthy’?

- popularity (‘likes’, downloads),
- use (existence of derived data sets or publications),
- Reproducibility/quality (is reproducing data possible (it isn’t for observations of dynamic systems)), what is the cost, NASA Data level),
- curation costs (e.g. storage size)
- importance (is the research relevant to important phenomena or societal issues?)
- has ‘minimal descriptive metadata’ available (e.g. domain vocabularies, dublin core)
Semantic Curation

- Traditionally, curator would really know the data and may not make decision criteria explicit
- Goal: Leverage what we know about the data to improve and scale curation by increasing curation automation and enabling the application of standardized policies across heterogeneous data collections:
  - Define an evidence of curation-worthiness
  - Map known metadata to evidence (per project/system)
  - semantic technologies are the right choice for putting the metadata and maps in machine-readable form
ACR Interfaces and Services

Zoomable Preview
Ingest workflows
DC:Creator link
Extracted metadata
License
Tags
Bulk Upload
Map Overlay of ACR Data
Note Emphasis on Water Resource SubDiscipline and Earth Sciences in General
Tupelo II: Semantic Content Management

- Web Protocol to
  - Authenticate
  - Get/Set Data
  - Get/Set Metadata

- POJO/Bean API

- Flexible Global Identifiers

- Secure Context Management

- Extensions to support specific ontologies (provenance, data streams, GIS, …)

- Extensions to support computational inference (spatial indexing, provenance, extracted metadata...)

Futrelle et. al.