Community Inventory of EarthCube Resources for Geoscience Interoperability

CINERGI Project Overview

ILYA ZASLAVSKY
RAQUEL CALDERON
CHRIS CONDIT
JEFFREY GRETHE
AMARNATH GUPTA
BURAK OZYURT
THOMAS WHITENACK
DAVID VALENTINE
ALICE GILIARINI
AARON GONG

STEPHEN RICHARD, LEAH MUSIL
Arizona Geological Survey

KERSTIN LEHNERT, LESLIE HSU
LDEO, Columbia University

TANU MALIK
University of Chicago

LUIS BERMUDEZ
Open Geospatial Consortium

University of California San Diego
CINERGI Goals

- Large inventory of high quality information resources across disciplines, with traceable provenance, usable across EarthCube research scenarios:
  - datasets, catalogs, vocabularies, information models, services, process models, repositories, etc.
- Engaging geoscience community in creating and curating the inventory
- Enabling search and integration across domains, and facilitating navigation across information objects of different types:
  - resources, people/organizations, publications, models, workflows, software, activities, etc.
Prior work

- HISCentral time series catalog and service registry
- EarthChem, PetDB, SedDB, SESAR, IGSN...
- NIF Resource Registry
- National Geothermal Data System
- NeuroScience Information framework, SciCrunch
- Cross-domain interop Roadmap
- Other relevant registry efforts, including GEOSS and multiple domain registries, catalog software and services, standardization efforts
Resource discovery dilemmas

- Not a high-level concern (not likely to attribute science results to discovery) – yet the most often cited need
  - What is missing in our understanding of discovery?
- What we discover:
  - Databases? Datasets? Observations? Features of interest?
  - Other types of resources? People?
- How we discover: Centralized vs distributed search architecture; iterative search and navigation
- What is result of discovery? When are you done with the discovery phase and decide to experiment with data?
Why we need inventories in EarthCube

- To answer question “What is EarthCube?”
- To see what is available, avoid duplication, identify gaps
- To support discovery, understanding, integration
- To organize communities around shared use of information resources
- To monitor resources and understand what is “reproducible”
- For getting credit for EC-certified resources
- To compose architectures from available resources
Metadata enhancing in CINERGI

Domain Inventories
RCN (Research Coordination Networks)

Domain workshops

CINERGI Metadata Pipeline

Catalogs

High-level assets
CINERGI metadata harvesting and content enhancement

**Harvest adapters**: description of information sources, allows connection and ingestion

**Staging database**: persist original harvested descriptions and updates from processing/curation

**Document processing components**: enhance content or presentation, update provenance record

**Public access components**: external interfaces to present content for users

---

CINERGI Community Inventory of Earthcube Resources for Geoscience Interoperability
## Components

<table>
<thead>
<tr>
<th>Resource Name</th>
<th>Brief Description</th>
<th>Interface</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>CINERGI UI search interface</td>
<td>Front-end web client for crossdomain search. Based on West Coast Ocean Data Portal (<a href="https://github.com/WCGA">https://github.com/WCGA</a>)</td>
<td>CSW/ISO 19139</td>
<td>Java/Geoportal/SOLR</td>
</tr>
<tr>
<td>CINERGI catalog search services</td>
<td>CSW, SOLR Rest, OpenSearch endpoint to access inventory metadata content</td>
<td>CSW/ISO 19139</td>
<td>Java/Geoportal/SOLR</td>
</tr>
<tr>
<td>CINERGI metadata enhancers</td>
<td>Services that update various elements of resource descriptions to add value, e.g. bounding boxes from place names, standardized keywords based on text analytics</td>
<td>REST/ ISO19139</td>
<td>Java</td>
</tr>
<tr>
<td>CINERGI ontology</td>
<td>Unified ontology for semantic tagging and discovery; assembled by harmonizing and integrating a variety of existing ontologies and vocabularies.</td>
<td>REST / SciGraph-RDF</td>
<td>Java</td>
</tr>
<tr>
<td>CINERGI Validation service</td>
<td>Services to validate metadata records in CINERGI; build on OGC TEAM engine technology.</td>
<td>REST/ ISO19139</td>
<td>Java</td>
</tr>
<tr>
<td>CINERGI Provenance service</td>
<td>Service to update and retrieve provenance information for metadata records</td>
<td>REST/W3C PROV RDF</td>
<td>Java</td>
</tr>
<tr>
<td>CINERGI provenance database</td>
<td>Database to store provenance info</td>
<td>REST/ JSON XML/PROV</td>
<td>Neo4J</td>
</tr>
<tr>
<td>CINERGI Community Resource Viewers</td>
<td>Applications to assist domain users in organizing and discovering domain resources</td>
<td>ODATA, Google Spreadsheet api / JSON</td>
<td>Silverlight/ HTML5</td>
</tr>
<tr>
<td>CINERGI metadata staging database</td>
<td>DB to store metadata documents</td>
<td>REST/JSON</td>
<td>MongoDB</td>
</tr>
</tbody>
</table>
Content enhancement components

- Common enhancer API
- Provenance recording: W3C PROV and Neo4J
- Spatial enhancer (bounding boxes)
- Keyword enhancer
  - Based on domain vocabularies; instruments; organizations; themes; platforms; time; resource types; repositories
  - Uses SciGraph API for semantic processing
- Validation and provenance components
Working with geoscience communities

Community inventory sources:

- End-user workshops
- High-level functional components
- EarthCube RCNs: “Collaboration and Cyberinfrastructure for Paleogeosciences (C4P)” and “Building a Sediment Experimentalist Network (SEN).” Demos [here](#)

- YOUR INVENTORY!

Inventory viewer/editor/faceted search online application
Get the stable isotopic measurements, from carbonates in North America of Silurian age, integrate them with all known fossil occurrences and geochronological estimates and rotate all of the data back to their proper paleocoordinates.

Identify common patterns in complex queries

Define a mechanism for extending CINERGI pipeline to accommodate complex queries:

- Registering and describing data resources
- Adding ontology describing the domain
- Specifying relationships between terms

Look for all GCMs that have at least a 10 degree resolution, contain a coupled Sea Surface Model, are written in Fortran, Unix compatible, and were created at NCAR, Boulder.

Look for all tree-ring analysis MATLAB software written by Dave Meko.
URLs to demo

- **Project web site:**
  - [http://earthcube.org/group/cinergi](http://earthcube.org/group/cinergi)

- **Search interfaces:**
  - [http://cinergi.cloudapp.net/](http://cinergi.cloudapp.net/) (after processing)

- **C4P viewers (there are other viewers, too)**
  - [http://hydro10.sdsc.edu/c4pviewer/CommunityPivot.html](http://hydro10.sdsc.edu/c4pviewer/CommunityPivot.html) (Silverlight)
  - [http://pivots.azurewebsites.net/c4p.html](http://pivots.azurewebsites.net/c4p.html) (new HTML5 app)

- **APIs:**
  - [https://docs.google.com/document/d/10UdfVN5zcj74deuW40YNnZyRZMVILk_AWV8QcF9PFd8/edit#heading=h.nj0h4bo3qjp3](https://docs.google.com/document/d/10UdfVN5zcj74deuW40YNnZyRZMVILk_AWV8QcF9PFd8/edit#heading=h.nj0h4bo3qjp3)
CINERGI’s role in EarthCube

If your data facility does manual metadata curation:
explore CINERGI pipeline and see if automatic metadata enhancement is useful; examine metadata provenance for your records, help us train the system

If you organize a domain community:
consider setting up and using a CINERGI community resource viewer

If you maintain a domain catalog:
consider interfacing it with CINERGI

Contribute to and help curate existing inventories, esp. high-level resources, functional components
this will be used in EC architecture development

Have interesting discovery use cases:
contribute use cases from your domain, see what we need to add to CINERGI to support them (e.g., additional vocabularies, data repositories, harvest adapters...)

CINERGI Community Inventory of Earthcube Resources for Geoscience Interoperability