

Metadata Enhancement in CINERGI

Ilya Zaslavsky
San Diego Supercomputer Center, UCSD

EarthCube

“EarthCube is a community effort to promote interdisciplinary geoscience by enabling technology, organization, and culture that facilitates connectivity through standards and protocols to existing and emerging resources.”

NSF & GEOSCIENCE COMMUNITY PARTNERSHIP

Geoscience community involvement in decision-making processes



EarthCube Strategy

Build on existing
resources, a SoS
approach

Architecture

Funded Projects

Develop CI that is
responsive to
community input
& assessment

Represent
community
and inform NSF

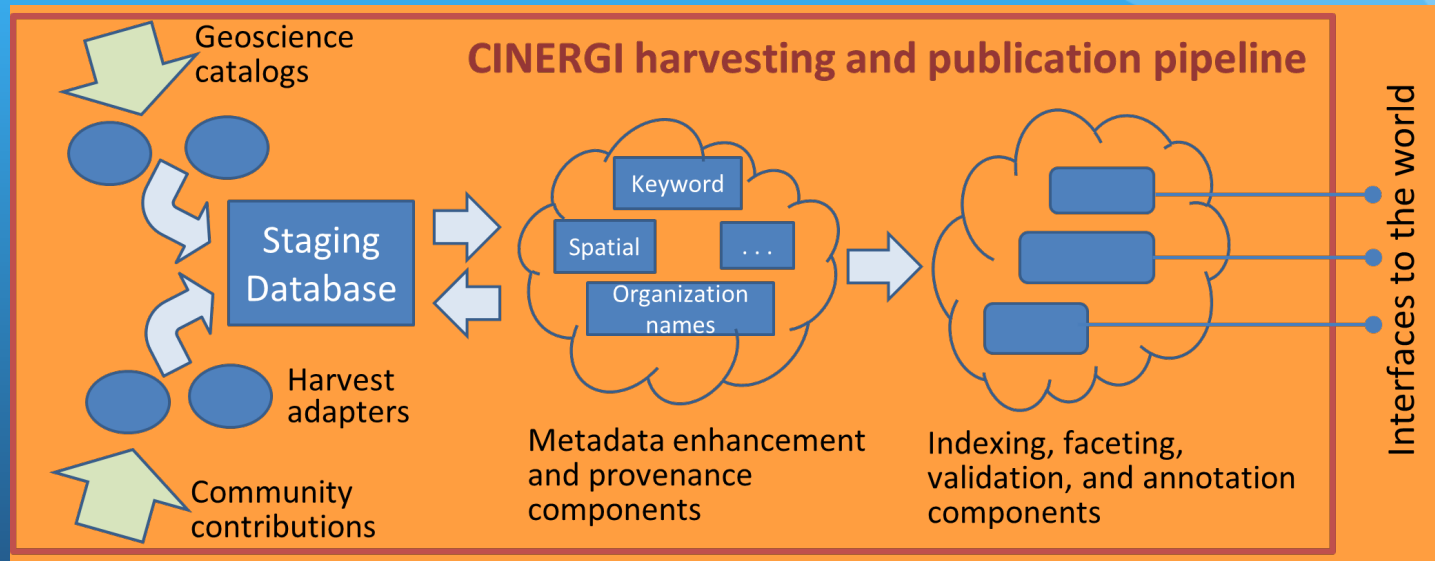
Governance

Community

Connect technical
development and
social/cultural
adoption

The CINERGI Metadata Pipeline

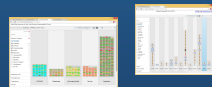
CINERGI: Community Inventory of EarthCube Resources for Geoscience Interoperability



Domain Inventories



Domain workshops



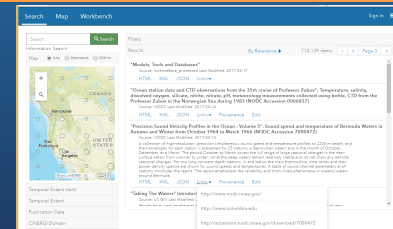
High-level assets



RCN
(Research
Coordination
Networks)



Content
enhancement



<http://cinergi.sdsc.edu/>

Content Enhancement Components

- ▶ Common enhancer API
- ▶ Provenance recording: W3C PROV and Neo4J
- ▶ Spatial enhancer (bounding boxes)
- ▶ Keyword enhancer
 - ▶ Materials; Processes; Equipment; Methods; Features; Activities; Science Domains; Geologic age; Organizations; Resource types
- ▶ Organization Enhancer
 - ▶ Associate with Virtual Authority Identifiers
- ▶ Collection Enhancer
 - ▶ Add keywords to a metadata collection
- ▶ Schema validation

The screenshot displays a web application interface with a map of the United States on the left, showing major cities like Vancouver, San Francisco, Los Angeles, and San Diego. Below the map is a 'Temporal Extent' section with a graph showing data points from 1500 to 2000. The right side of the interface shows a list of publications with their titles and abstracts. The first publication is titled '"Basin scale" versus "localized" pore pressure/stress coupling - Implications for trap integrity evaluation' and the second is '"Chill Out" Oregon Institute of Technology is a Winner'.

Temporal Extent (text)

Temporal Extent

Publication Date

CINERGI Domain

CINERGI Material

- ▶ material (other) (101979)
- ▶ environmental material (92582)
 - ▶ water (51405)
 - ▶ mineral (18816)
 - ▶ groundwater (12026)
 - ▶ petroleum (10337)
 - ▶ surface water (2553)
 - ▶ soil (2401)

"Basin scale" versus "localized" pore pressure/stress coupling - Implications for trap integrity evaluation

Source: ScienceBase_processed Last Modified: 2017-03-18

HTML XML JSON Links Provenance Edit

"Chill Out" Oregon Institute of Technology is a Winner

Source: US GIN Last Modified: 2017-03-15

HTML XML JSON Links Provenance Edit

"D" Ore Body Coronado Copper and Zinc Company

Source: US GIN Last Modified: 2017-03-15

ADMRR map collection: "D" Ore Body Coronado Copper and Zinc Company; 11 x 8 in.

HTML XML JSON Links Provenance Edit

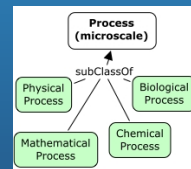
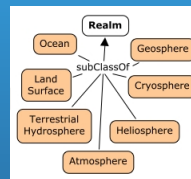
"Das Operationsziel ist Stabilität": Klinische Untersuchung zur Stabilität des Daumen-Grundgelenkes bei Kind

GeoSciGraph and Ontologies

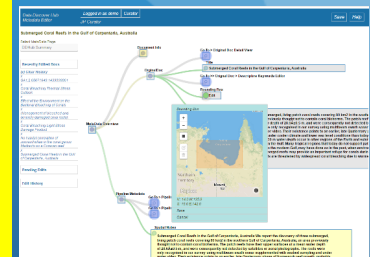
GeoSciGraph ontology management system provides semantic infrastructure. It relies on a cross-domain ontology of geoscience terms, amalgamating several independently developed ontologies or taxonomies

Some included ontologies:

- SWEET
- ENVO
- CHEBI
- YAGO (geo features)
- NASA GCMD (equipment, providers)
- GeoSciML
- Geochronology
- EDAM Bioinformatics (software terms and operations)
- Also: VIAF



Added annotation properties for combining ontology fragments (*cinergiFacet*, *cinergiParent*)

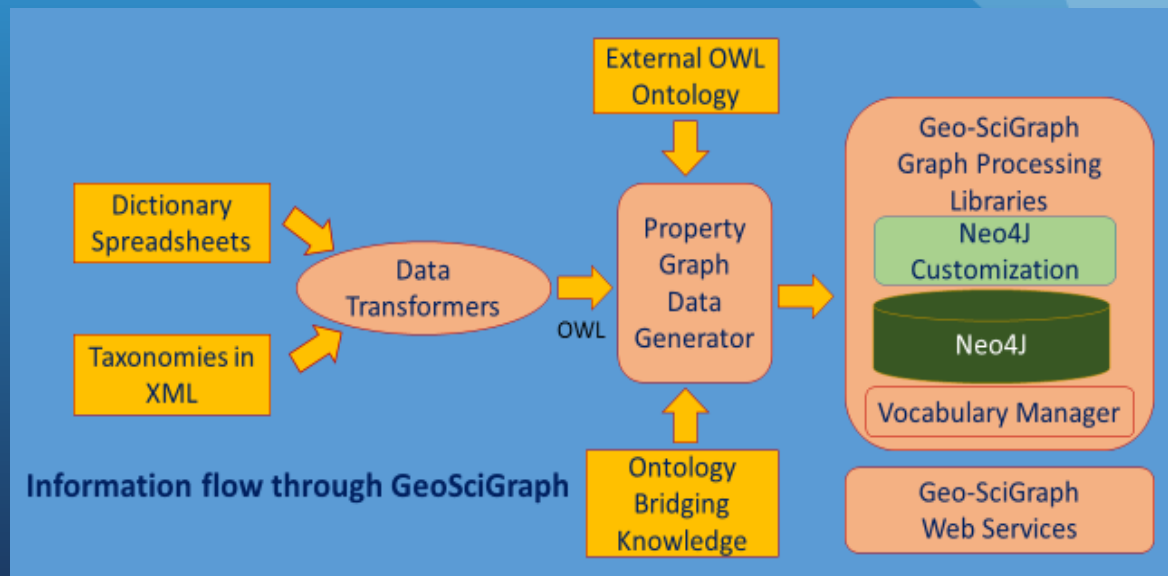


Metadata editor
Approve or discard semantic annotations



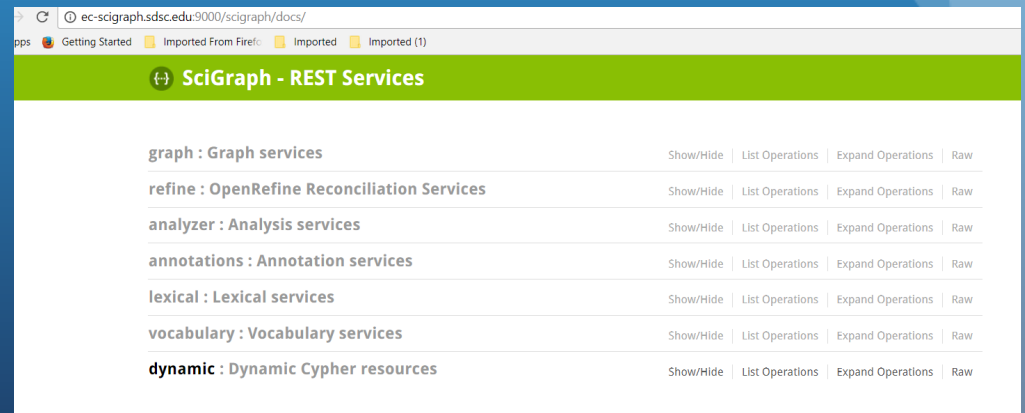
GeoSciGraph Services API

- GeoSciGraph Services: The GeoSciGraph API exposes a set of web services for querying and exploring the CINERGI ontology.
- Lexical Services are used to break text into sentences and perform sentence parsing using lightweight NLP techniques.
- Vocabulary Services are used to find concepts, synonyms, term categories, autocomplete search, and term suggestions based on similarity.



GeoSciGraph Services API

- **Graph Services** are used to navigate the graph by following user-specified relationships and finding neighborhoods. Another service locates the head of a clique (all pair connected subgraph) in an ontology graph.
- **Refine Services** provides a gateway to OpenRefine, Google service to match entries in a data table to an ontology.
- **Cypher Utility Service** is a pass-through service that directs a user-specified Cypher query directly to the underlying Neo4J system.
- **Analyze Services** provides a way to add custom-defined analyses into the GeoSciGraph system



The screenshot shows a web browser window with the URL `ec-scigraph.sdsc.edu:9000/scigraph/docs/`. The page title is "SciGraph - REST Services". Below the title, there is a table listing various services and their corresponding actions.

Service Name	Show/Hide	List Operations	Expand Operations	Raw
graph : Graph services	Show/Hide	List Operations	Expand Operations	Raw
refine : OpenRefine Reconciliation Services	Show/Hide	List Operations	Expand Operations	Raw
analyzer : Analysis services	Show/Hide	List Operations	Expand Operations	Raw
annotations : Annotation services	Show/Hide	List Operations	Expand Operations	Raw
lexical : Lexical services	Show/Hide	List Operations	Expand Operations	Raw
vocabulary : Vocabulary services	Show/Hide	List Operations	Expand Operations	Raw
dynamic : Dynamic Cypher resources	Show/Hide	List Operations	Expand Operations	Raw

The screenshot displays a web application interface with a search bar at the top left and a 'Workbench' tab. Below the search bar, there are filters and a list of search results. The first result is titled '*Models, Tools and Databases*' with a source of 'ScienceBase_processed' and a last modified date of '2017-03-17'. The second result is '*Ocean station data and CTD observations from the 35th cruise of Professor Zubov*', with a source of 'NODC' and a last modified date of '2017-03-14'. The third result is '*Precision Sound Velocity Profiles in the Ocean - Volume 5*', with a source of 'NODC' and a last modified date of '2017-03-14'. On the right side, there is a map showing the United States and Canada, with markers for Vancouver, San Francisco, Los Angeles, and Miami. Below the map, there are options for 'Temporal Extent (text)', 'Temporal Extent', 'Publication Date', and 'CINERGI Domain'. The main content area shows a document titled '*Submerged Coral Reefs in the Gulf of Carpentaria, Australia*' with a 'Document Info' section, a 'Metadata Overview' section, and a 'Bibliography' section. The 'Bibliography' section includes a citation: 'Submerged Coral Reefs in the Gulf of Carpentaria, Australia We report the discovery of three submerged, living patch coral reefs covering 88 km² in the southern Gulf of Carpentaria, Australia, an area previously thought to be devoid of coral reefs. The patch reefs have their upper surfaces at a mean water depth of 28-32 m, and were subsequently not detected by satellite or aerial photography. The reefs were only recognized on our 1st survey using multibeam swath sonar supplemented with seabed sampling and under water video. These shallow reefs are an exciting new discovery for coral reef science and are threatened by widespread coral bleaching due to warmer temperatures.' The interface also includes a 'Map' section with a 'Bounding Box' and a 'Save' button.

Over 1 million records

From multiple repositories and EC contributions

Metadata automatically enhanced through CINERGI

- Provenance tracing
- Metadata editing
- Automated semantic processing and indexing
- Faceted search
- Integration with EC Workbench

Metadata Sources in DDH

USGS

NOAA/NCEI

CUAHSI

Data.gov

USGIN

NGDS

CZO

NCDC

CORIS

DataCite

Re3data/databib

EarthCube projects

RCNs: C4P, SEN, ECOGEO, CRESCYNT

Model catalogs (NOAA, EPA, TESS)

Geoscience Australia

OpenTopography

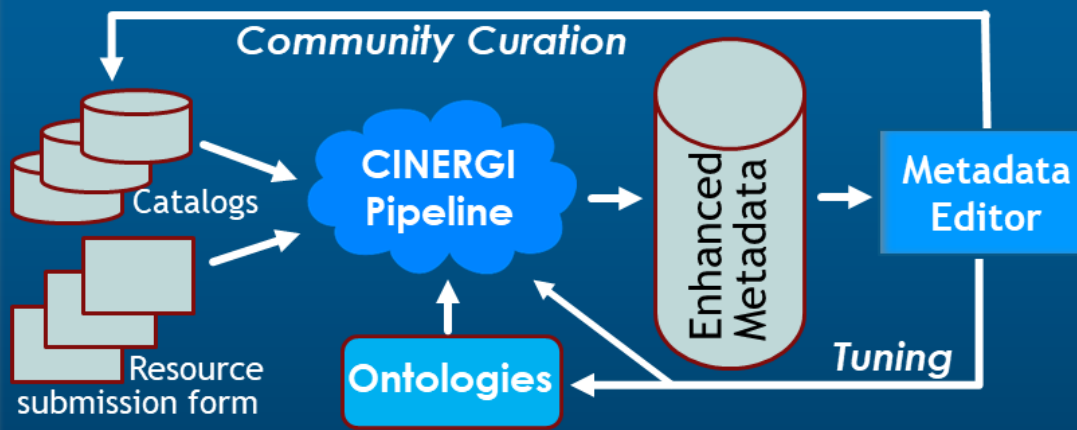
IEDA (MGDS, ECL, MetPetDB)

EarthRef MagIC

CINERGI Cyberinfrastructure resources

DDH use cases

New Curation Model



Contribute A Resource

Resource Type:
Dataset

Resource Title:

Description / Abstract:

Pending Edits
Edit History

Logged in as demo Curator
JP Curator

Gulf of Carpentaria, Australia

Document Info

OriginalDoc

Go To > Original Doc Detail View

Title

Submerged Coral Reefs in the Gulf of Carpentaria, Australia

Go To > Original Doc > Descriptive Keywords Editor

Bounding Box

Edit

Metadata Overview

Pipeline Metadata

Go To > Pipeline

Go To > Pipeline

Save

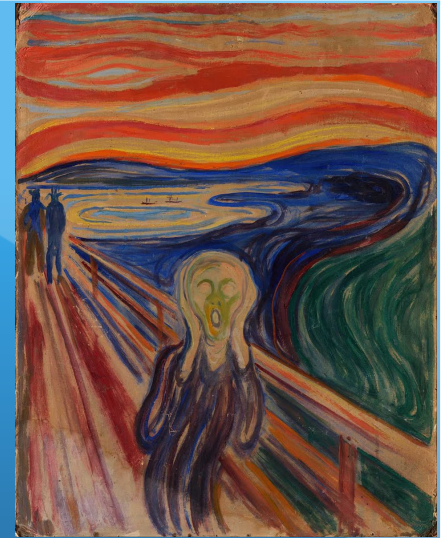
Cancel

Spatial Notes

Submerged Coral Reefs in the Gulf of Carpentaria, Australia We report the discovery of three submerged, living patch coral reefs covering 80 km² in the southern Gulf of Carpentaria, Australia, an area previously thought not to contain coral bioherms. The patch reefs have their upper surfaces at a mean water depth of 28.6±0.5 m, and were consequently not detected by satellites or aerial photographs. The reefs were only recognised in our survey using multibeam swath sonar supplemented with seabed sampling and under water video. Their existence points to an earlier, late Quaternary phase of framework reef growth, probably

Interesting issues...

- Semantic assignment
 - Selecting which ontology IDs to use when conflicts (machine learning)
 - Our ability to detect concepts and assign keywords may not match ontology's level of detail
 - Scalable ontology alignment
- Enabling faceting and search
 - Pre-defining upper facets; adjusting underlying ontology fragments for consistency; customizing for specific communities (eg promoting domain-specific search facets)
- Generating corpus of text to analyze
- Cross-granularity search
- Criteria of success (Quality of search? Interoperability? Engaging domain users)



DDH is based on CINERGI technology for:

- Metadata enhancement, using an extendable metadata augmentation platform
- Semantic annotation, leveraging text analytics and a large composite ontology
- Distributed metadata curation
- Faceted search

DDH focus:

- Improved search over expanded inventory
- Enhanced ontology: additional vocabularies and relationships (semantic proxy, spatial, temporal)
- Deeper dataset and repository registration
- Search across granularity levels
- EarthCube workbench integration
- Complex discovery use cases

Workshop in the afternoon

- Four key operations:
 - Search (text, spatial, temporal), over generated facets
 - Creating and indexing a semantically-enhanced ISO document
 - Metadata editing
 - Adding your own enhancer
- Exploring the ontology
- Exploring service APIs
 - SciGraph
 - CINERGI/foundry