SEMANTICS AND LINKED OPEN DATA: THE OCEANLINK PROJECT

Tom Narock

Marymount University

tnarock@marymount.edu

INTRODUCTION TO OCEANLINK

- A wide spectrum of maturing methods and tools, collectively characterized as the Semantic Web, is helping to vastly improve the dissemination of scientific research.

- Creating semantic integration requires input from both domain and cyberinfrastructure scientists.

- The OceanLink project, an EarthCube Building Block, is demonstrating semantic technologies through the integration of ocean science data repositories, library holdings, conference abstracts, and funded research awards.

OCEANLINK TEAM

- *Lamont-Doherty Earth Observatory*
  - *Robert Arko*
  - *Suzzanne Carbotte*
- *Marymount University*
  - *Tom Narock*
- *University of Maryland, Baltimore County*
  - *Tim Finin*
- *Woods Hole Oceanographic Institution*
  - *Cynthia Chandler*
  - *Adam Sheperd*
  - *Peter Wiebe*
  - *Lisa Raymond*
- *Wright State University*
  - *Pascal Hitzler*
  - *Adila Krisnadhi*
  - *Michelle Cheatham*

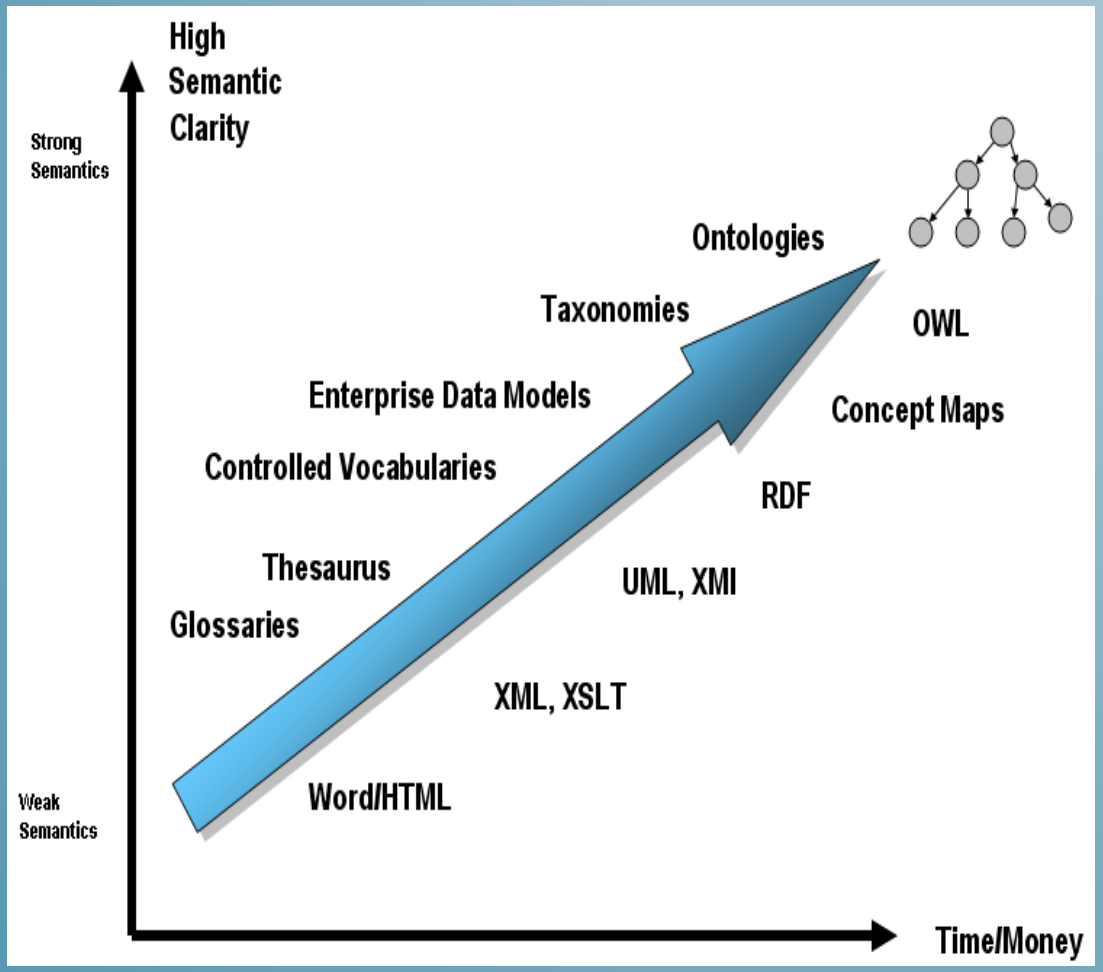
EARTHCUBE END-USER WORKSHOPS

Current challenges to interdisciplinary science:

- Cyberinfrastructure for Paleogeoscience, Feb. 4-6, 2013
  - "Unawareness and/or underutilization of standards for data/metadata"
  - "Difficulty of importing/exporting data from databases"

- Ocean 'Omics Workshop, August 20-23, 2013
  - "The ocean 'omics community would benefit from "Google-like" search and suggestion functions/engines, that could query across complex and heterogeneous, federated environments"
  - "The community would benefit from access to a web clearing house/portal with links to standard … best practices, algorithms, software and workflows …"

SPECTRUM OF SEMANTIC WEB TECHNOLOGIES



- OceanLink seeks strong semantics
- This must be balanced with communities ability to implement and use
- Part of our cyberinfrastructure is an attempt to lessen this burden

Image Credit: http://www.mkbergman.com/936/brown-bag-lunch-an-intrepid-guide-to-ontologies/

WHAT IS LINKED OPEN DATA

A data publication methodology adhering to four rules

1. Use unique identifiers for things
2. Use HTTP and HTTP URIs so that things are web accessible
3. An identifier should return semantic statements for machine and human consumption (RDF)
4. Include links to other related identifiers

WHAT DATA ARE WE LINKING?

- We want to link data to related data as well as to publications, presentations, funded awards, and gazetteers
- Essentially anything that provides context and may also be relevant to use

Rolling Deck to Repository, central repository for research vessels

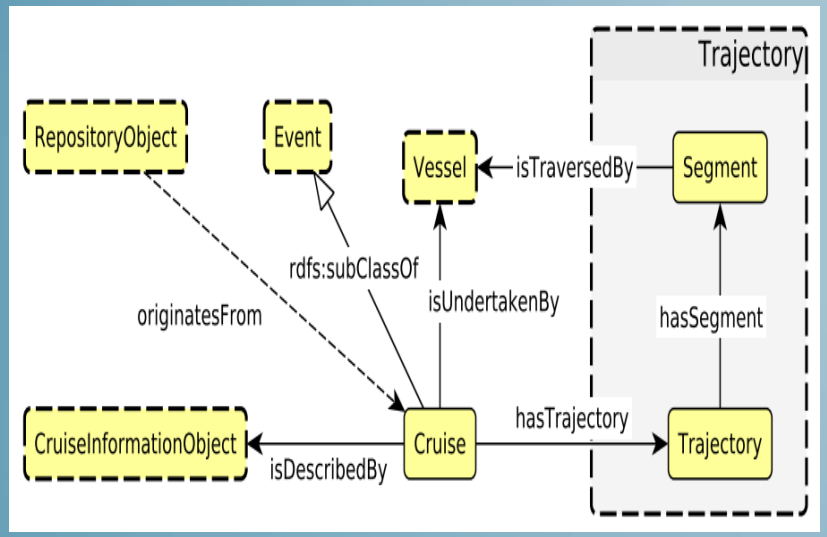
Biological and Chemical Ocean Data Archive

Cruise reports and PhD theses

All Funded Awards 1976 thru present

Conference Presentations 1995 thru present

SEMANTIC PATTERNS

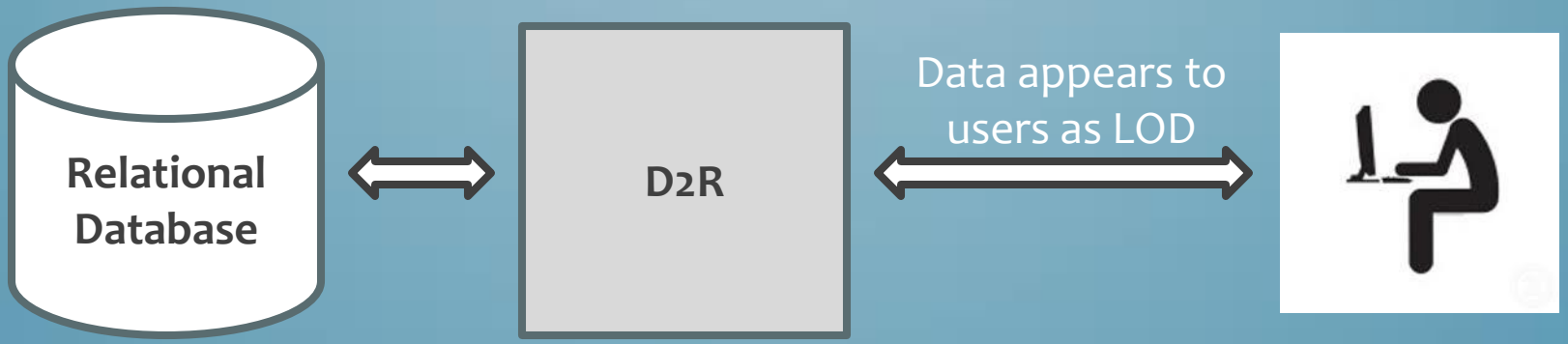


- Basic building block is the ontology
- How to provide large scale integration?
  - Upper Ontology
  - Ontology Design Patterns
- We have chosen Ontology Design Patterns
- Patterns
  - Agent (Person, Organization)
  - Award
  - Cruise
  - Repository
  - Publication
- Don’t need to implement all patterns
- Some specific to oceans, many reusable

MULTIPLE WAYS TO IMPLEMENT

- Can republish metadata with LOD tools and methodologies

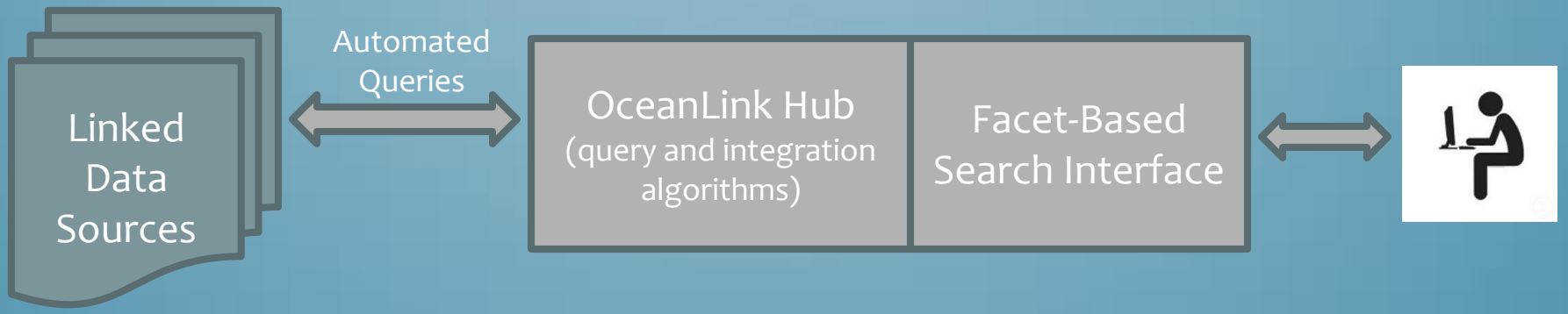
- There also exist tools to map existing relational databases to LOD



CYBERINFRASTRUCTURE

1. Querying multiple LOD sources can be challenging
2. LOD is more useful when data is interlinked

- We are building a "hub" to
  - Automatically identify links between datasets (e.g. cruises in AGU/NSF abstracts)
  - Co-entity resolution
  - Provide a user friendly facet-based search interface



OCEANLINK PORTAL

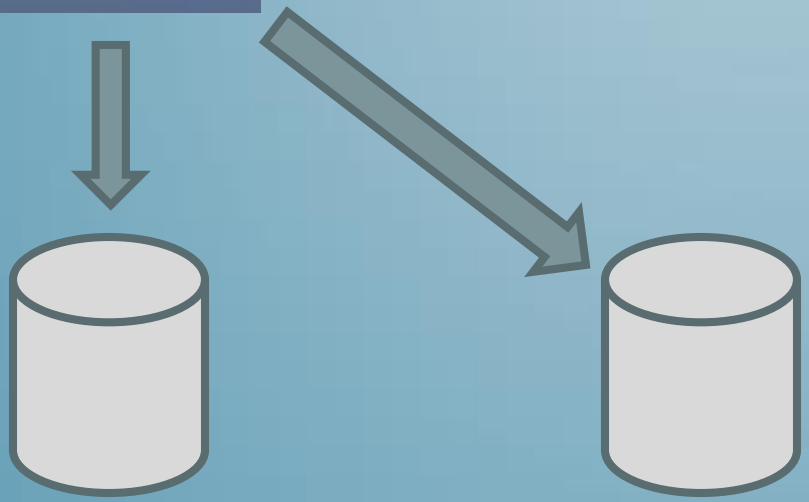
- Facet-based search interface
- Initial version will support searching via
  - Cruise
  - Agent (Person or Organization)
  - Funding
  - Program
- Results will
  - Point users to data sources
  - Highlight links between repositories
  - Identify supporting materials such as publications and grants

INTERFACE EXAMPLE

Research cruise vessel: Maurice Ewing
Operator: Lamont-Doherty
Cruise ID: EW0408

INTERFACE EXAMPLE

Research cruise vessel: Maurice Ewing
Operator: Lamont-Doherty
Cruise ID: EW0408



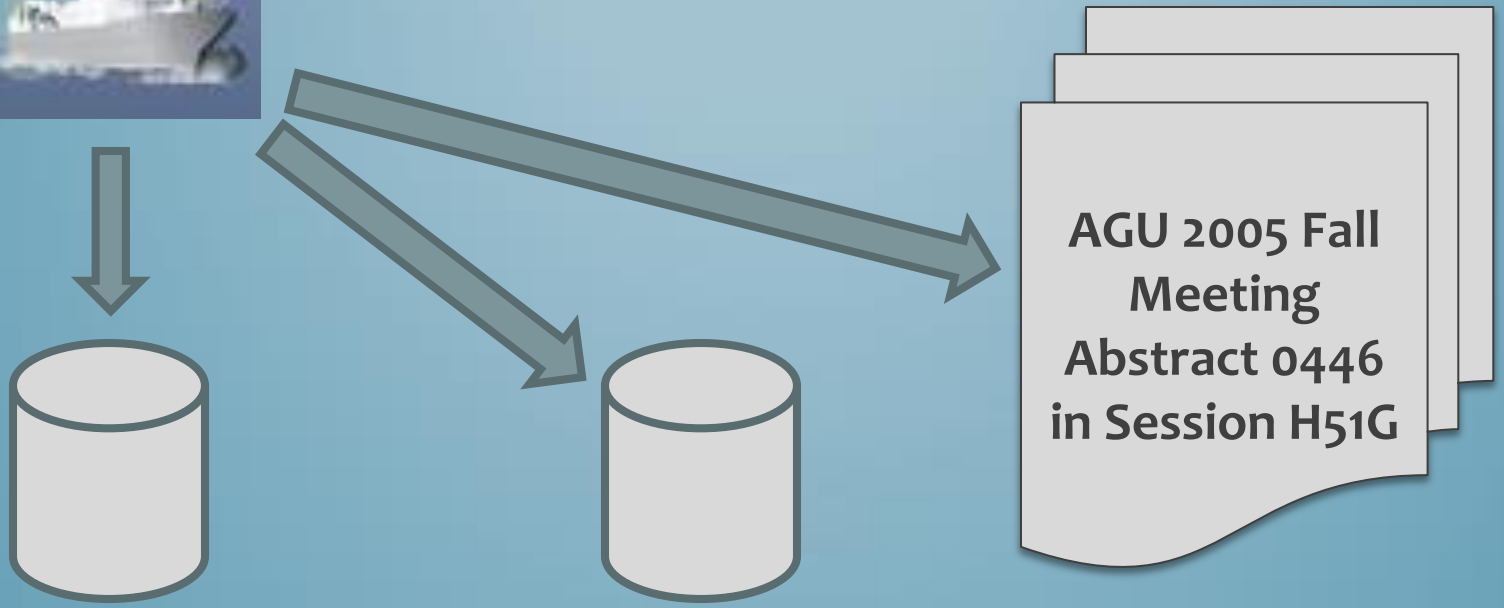
Cruise Logs and Cruise Data available at R2R (Columbia University)

This particular cruise no data at BCO-DMO, but this vessel has supplied data

Semantics allow the system to "know" that vessels at R2R and BCO-DMO are the same

INTERFACE EXAMPLE

Research cruise vessel: Maurice Ewing
Operator: Lamont-Doherty
Cruise ID: EW0408



Cruise Logs and Cruise Data available at R2R (Columbia University)

This particular cruise no data at BCO-DMO, but this vessel has supplied data

INTERFACE EXAMPLE

Part of the power of Linked Open Data is that users can follow links to explore relationships

- What other related data is available?
- What other publications did author create?
- What other funded projects mention my data?

INFERENCE AND REASONING

- Ontologies are formal semantic models
- If, for example, data is asserted to be of type “Cruise” then this implies certain properties and relationships
- Computers can use tools called reasoners to infer what else must be true give certain assertions
- For example, if an instrument is said to be onboard a cruise then that cruise can be inferred to collect data of a specific type – even though not explicitly stated

QUALITY ASSURANCE

- A primary use of reasoning will be in Quality Assurance

- Rolling Deck to Repository states that a cruise ran from 1/2013 to 6/2013
- BCO-DCMO states that same cruise ran from 1/2012 to 6/2012
- OceanLink ontology indicates that a cruise may only have one start and one stop date

- Semantic Web tools reason that data is inconsistent
- This type of Quality Assurance can be done by individual data providers or by the OceanLink hub
- Use of semantics and standards means common tools

CROWD SOURCING

- Data is messy and not all of the links we infer will be valid
  - AGU: T. Narock, T. W. Narock
  - NSF: Thomas Narock
- We'd like to solicit our users to help validate our links
- Use Social media on results page
- Exploring ResearcherID and Orcid ID

ORCID
Connecting Research and Researchers

RESEARCHERID

RELEVANCE TO C4P

Overlap in data – primarily in the cruises. Many cruises collect sample data (physical specimens like rocks, sediments, fluids, etc) as well as environmental sensor data.

Semantic Web approaches, and Linked Data specifically, are a nice way to connect the samples with other data (both field data and post-field analysis/publications) related to a particular cruise.

Globally unique identifiers are needed in many C4P applications (and outside C4P). Linked Data requires unique identifiers as part of the technology

OceanLink and C4P share many common controlled vocabularies for science themes, devices, data types, gazetteers (seafloor features), etc.

OceanLink is gaining practical experience both in the technologies as well as the socio-techncial aspects

TIMELINE

Completed
- Meetings with stake holders to create ontology design patterns

Ongoing
- Have some datasets available via Linked Open Data
- Remaining datasets actively being deployed as Linked Open Data
- Search interface is currently being developed

Future
- Branch out to related areas of geoscience, e.g. ecology and paleo

Expectations
- First version of OceanLink hub to be ready by EarthCube meeting this summer
- Several of us will be attending C4P meeting this summer

http://www.oceanlink.org
http://schema.oceanlink.org