Vocabularies, Ontologies and Ocean Biogeochemistry Data "*Matching and Mapping*"

Cyndy Chandler

Biological and Chemical Oceanography Data Management Office
BCO-DMO

24 January 2011

Marine Biological Laboratory
Woods Hole, MA USA

Acknowledgments

- Dicky Allison, Bob Groman and Tobias Work (BCO-DMO, Woods Hole Oceanographic Institution)
- Patrick West, Stephan Zednik and Peter Fox (Tetherless World Constellation, Rensselaer Polytechnic Institute)
- Andy Maffei (Ocean Informatics Working Group, WHOI)

Discussion Topics

- background
- goals and benefits? why do this?
- vocabularies and ontologies
- BCO-DMO vocabularies
- BCO-DMO ontology
- what's next?


Woods Hole, Massachusetts, USA

Where did we come from? some background . . .

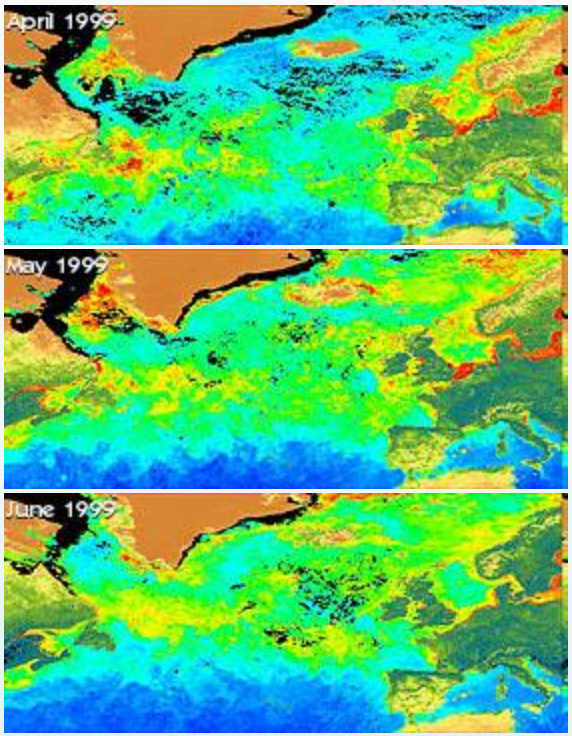
- BCO-DMO evolved from separate data offices for US JGOFS and US GLOBEC
  US JGOFS = US Joint Global Ocean Flux Study
  US GLOBEC = US GLOBal ocean ECosystems dynamics
- BCO-DMO = US JGOFS + US GLOBEC + others
  funded in late 2006 by NSF OCE
- the US JGOFS and US GLOBEC data systems were designed to serve their respective programs
- independent but similar data systems
- separate, but similar and overlapping vocabularies to describe the data and metadata

Where are we going? the goal . . .

- data managers want to make it easier to
  - locate data of interest
  - assess ‘fitness for purpose’
  - integrate data from distributed sources
- the ultimate objective of the vocabulary mapping and ontology project is to improve data discovery, access, and integration and provide support for unambiguous, machine-interpretable data resources

“Scientists are confronted with significant data management problems due to the large volume and high complexity of scientific data. In particular, the latter makes data integration a significant technical challenge.” (A.K. Sinha, Geoinformatics ‘Data to Knowledge’, 2006)

Why do this? Why now?



New research paradigms . . .

- science themes trending toward
  - interdisciplinary
  - basin-wide
- studies involving coupling of complex models
  - atmospheric and hydrologic
  - end-to-end food web

. . . require access to data from many disciplines

New expectations for data access

- complex research themes (ocean biogeochemistry, marine ecosystem research) require access to data collected by other researchers

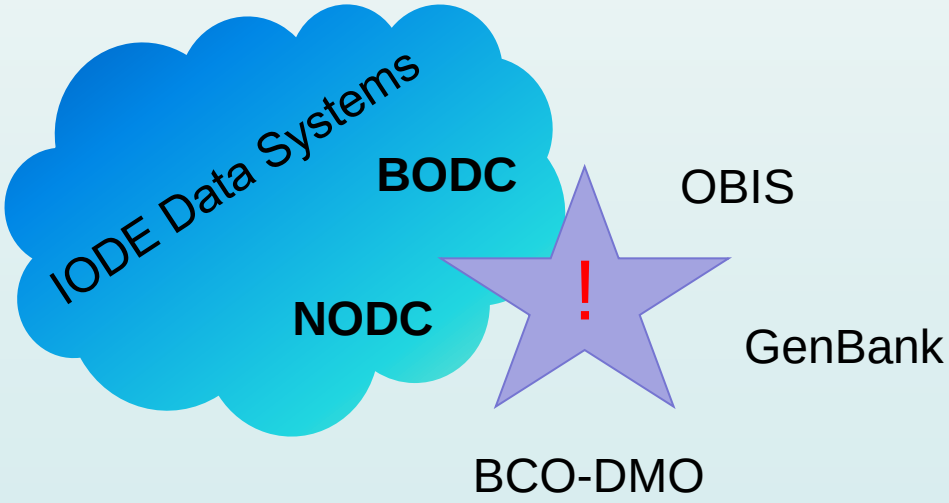
- access to results to enable science-based decision support for legislative policies
  - social science
  - economics
  - history
  - broad range of disciplines

New expectations for data access

- New tools based on emerging technologies are being developed to address the challenge of

  integration of distributed heterogeneous data

- informatics
- semantic mediation
- controlled vocabularies
- registered ontologies
- federated search
- faceted browse



New expectations for data access

- ocean science data accompanied by human readable metadata are of great value
- many of the new technologies assume that data resources will be accompanied by machine-readable metadata
- metadata paired with semantic encoding can enable more powerful interface tools for data discovery and access

Review . . .

- Where did we come from?
  independent but similar data systems with human-readable flat files of metadata

- Where are we going?
  improved data discovery, access, and integration; unambiguous, machine-interpretable data resources to support new research paradigms

- How do we get there?

controlled vocabularies

- a controlled vocabulary is a restricted list of defined terms (as opposed to free text)

- parameter names (measurements)
  e.g. nitrate, nitrate low level, nitrate plus nitrite
- Instrument names (sampling and analytical devices)
  e.g. CTD, Niskin bottle, MOCNESS

- the terms have definitions, e.g. a description of the sampling or analytical device (ideally this includes at least type, make and model)

ontologies

- an ontology is an explicit formal definition of terms and concepts and their relationships
- an ontology can encode the knowledge that expresses relationships between terms in the controlled vocabularies, and related groups of terms (classes), forming the basis for basis for inference and automated reasoning

- Together, controlled vocabularies and ontologies yield machine readable (and unambiguous) definitions of information stored in the database.

- Want to know more about ontologies?
  Noy, Natalya F. and McGuinness, Deborah L. (2001) Ontology Development 101: A Guide to Creating Your First Ontology, Knowledge Systems, AI Laboratory, Stanford University (KSL-01-05)

BCO-DMO challenge

- How do we provide semantically-enabled interfaces to legacy databases of heterogeneous, research data?

- Today the BCO-DMO database includes data from 16 programs and 135 projects, 1000 contributing PIs, 800 cruises, and over 1100 types of measurements from 210 different instrument types, with more data contributed daily.

- Investigators want to search the database for data of interest, but they use search terms that are different from those recorded in the database. e.g. a request for pigment data should find chlorophyll (all types) data sets.

the process . . . vocabularies . . .

1. organize the terms
   migrated our metadata from human-readable flat files to a relational database

2. created controlled vocabulary lists for many of the fields by “cleaning up” the term lists and using those to control data input; resolving duplicate, misnamed and ambiguous terms, for example:
   - people names
   - ship/vessel names (mapped to ICES codes)
   - cruise identifiers (authoritative cruise IDs)
   - instruments: sampling and analytical devices
   - measurements (data)

the process . . . continued . . .

3. In the BCO-DMO database, we want to have the values (instances) in as many fields as possible, be the terms from a controlled vocabulary (minimal free text).

Advantages of using controlled vocabularies:

1. aids data entry
2. enables mapping to other resources
3. facilitates interoperability
4. improves efficacy of search interface

Instrument Controlled Vocabulary

1. review and ‘clean up’ the list
   limited number of terms, no duplicates, no ambiguity

2. matching and mapping
   1. contributed name matched and mapped to the local BCO-DMO controlled vocabulary name
   2. BCO-DMO instrument name matched and mapped to the most specific term in the SeaDataNet Device Categories vocabulary

- http://vocab.ndg.nerc.ac.uk/list/L05/current
  served from the British Oceanographic Data Centre (BODC) Natural Environment Research Council (NERC) Data Grid vocabulary server
  http://www.bodc.ac.uk/products/web_services/vocab/

Example

- dataset contributed to BCO-DMO with instrument identified as SIO-CTD (this is an instance)

- matched and mapped locally within the BCO-DMO database to CTD Sea-Bird 911

- the BCO-DMO “CTD Sea-Bird 911” instrument is then matched and mapped to the SeaDataNet term Sea-Bird SBE 911 CTD

  using the SeaDataNet dereferencable URL

- http://vocab.ndg.nerc.ac.uk/term/L221/16/TOOL0035

SeaDataNet vocabulary server

- http://vocab.ndg.nerc.ac.uk/term/L221/16/TOOL0035
  <skos:Concept rdf:about='http://vocab.ndg.nerc.ac.uk/term/L221/16/TOOL0035'>
  <skos:externalID>SDN:L221:16:TOOL0035</skos:externalID>
  <skos:prefLabel>Sea-Bird SBE 911 CTD</skos:prefLabel>
  <skos:altLabel>SBE 911 CTD</skos:altLabel>
  <skos:definition>High precision and accuracy CTD made up from a Sea-Bird SBE 9 underwater unit and a SBE 11 deck unit. The underwater unit comprises protective cage (usually with a rosette) holding a pressure unit and temperature/conductivity unit. The latter is connected to a pump-fed plastic tubing circuit that may include other sensors. All plumbed and non-plumbed instruments (e.g. transmissometers and light meters) on the package are logged by the SBE 11. The unit was replaces by the SBE 911plus in 1997.</skos:definition>

Instrument Mapping Examples

original data set	BCO-DMO	SeaDataNet
SIO-CTD	CTD Sea-Bird 911	Sea-Bird SBE 911 CTD /term//L221/16/TOOL0035
bongo tow	bongo net	plankton net /term/L051/4/22
XBT	Expendable Bathythermograph	bathythermographs /term/L054/7/132

local → global

disambiguation of terms and concepts →

matching and mapping . . .

- requires time
  - do simpler terms first
  - more difficult terms later
- requires knowledge of the terms
  - may need additional team members
  - access to authoritative sources of information
  - additional research required
- tools can help
  - Excel, Google refine (Alex Dorsk, 10 Jan 2011)

Local and global term mapping

- the local term mapping within the BCO-DMO database allows investigators to continue using familiar names
- BCO-DMO controlled vocabulary use helps with data entry and database maintenance
- BCO-DMO instrument and parameter vocabularies include classes and subclasses (non-hierarchical)
- matching and mapping to a community vocabulary improves data discovery and data interpretation by reducing ambiguity of terms

Matching and Mapping Strategy

- For each dataset in the BCO-DMO database, we identify an instrument (sampling device, sensor) used in acquisition
- and match it to the most specific term in the BCO-DMO instrument controlled vocabulary list

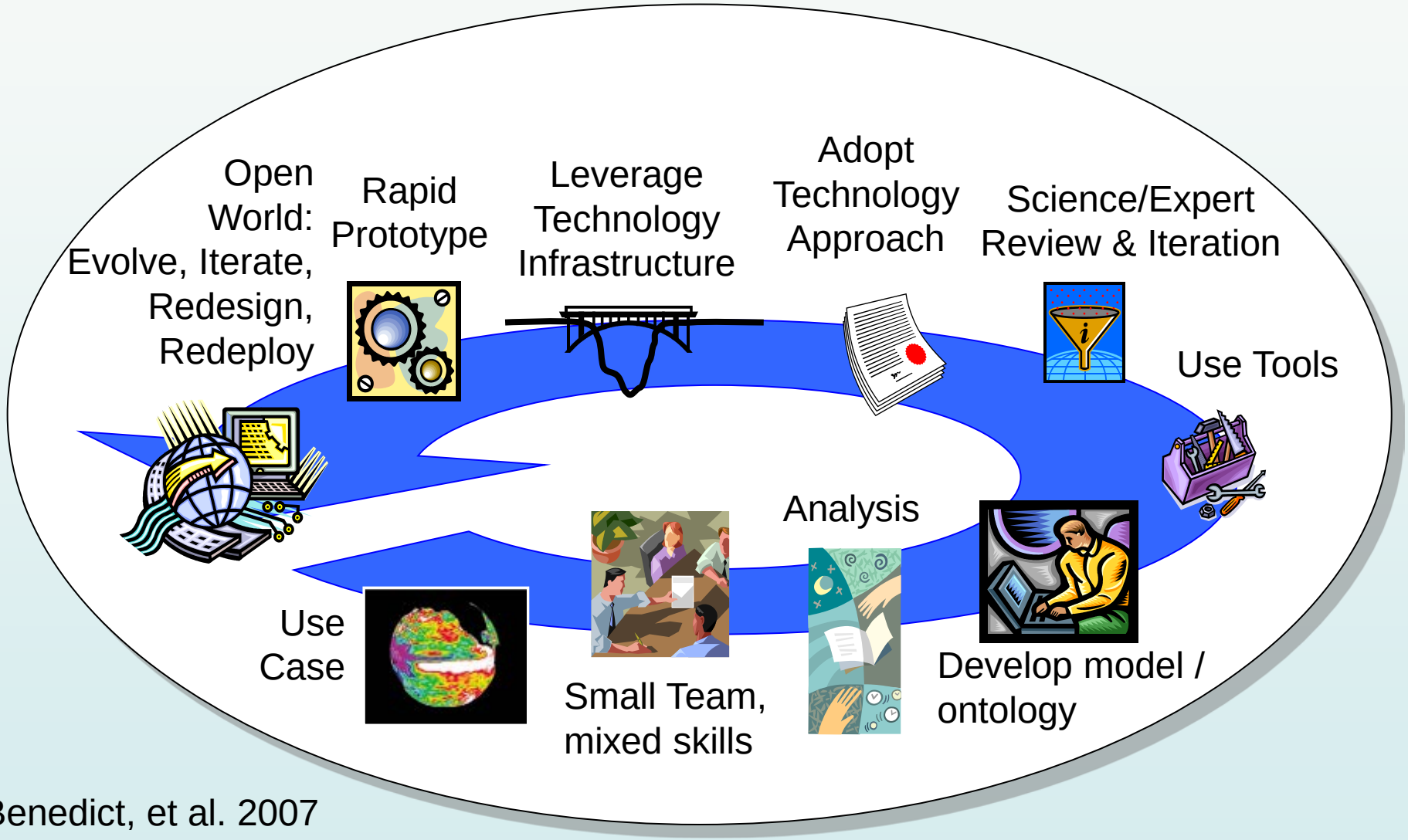
- As many instruments as possible (instrument name with type, make and model or instrument class) in the BCO-DMO instrument list are mapped to device terms in the SeaDataNet vocabulary – to the most specific term that is accurate
- When mapped, it inherits the relationships (explicit and implicit) of the term to which it is mapped.

Choosing a Vocabulary for Mapping

- The SeaDataNet device categories vocabulary and related sampling and sensor type vocabularies met our criteria for a community standard vocabulary:
- availability (Web-accessible via HTTP URIs)
- quality (completeness, clarity and precision, relevance)
- community adoption
- effective governance structure

Graybeal, J. 2009. "Choosing and Implementing Established Controlled Vocabularies."
In The MMI Guides:
Navigating the World of Marine Metadata.
http://marinemetadata.org/guides/vocabs/cvchooseimplement

Semantic Web Methodology and Technology Development Process



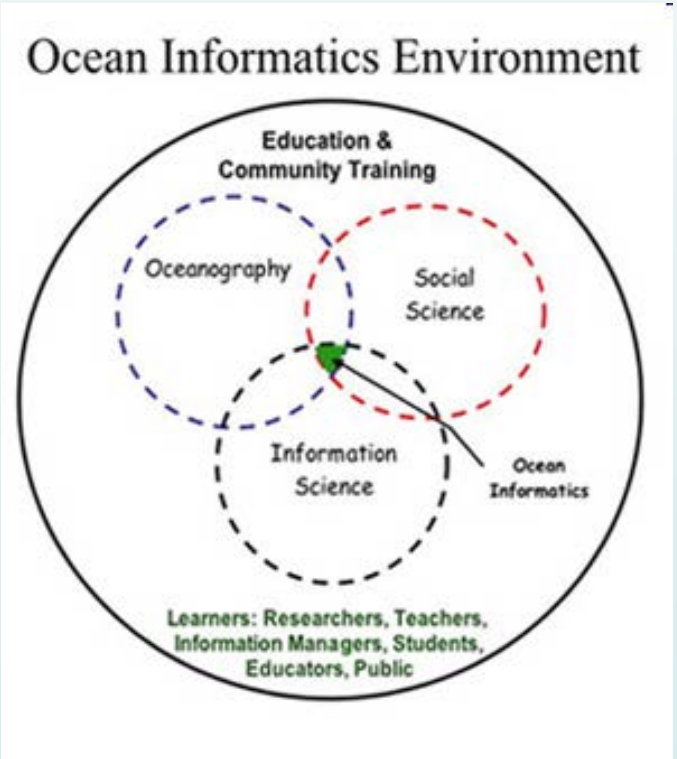
Benedict, et al. 2007
Eos, AGU IN53A-0950

Informatics Approach

Ocean Informatics is the union of oceanography, information science and social science domains.

Karen Baker (SIO)

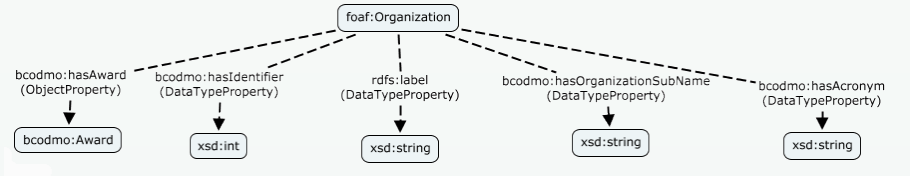
(DSR II, Baker & Chandler, 2008)
http://oceaninformatics.ucsd.edu/

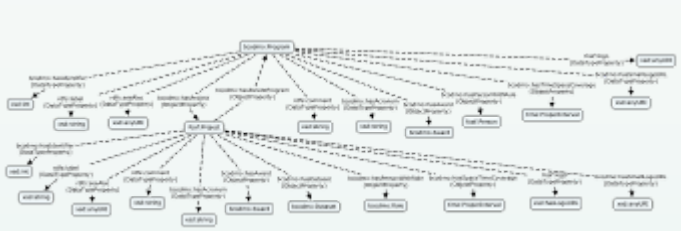


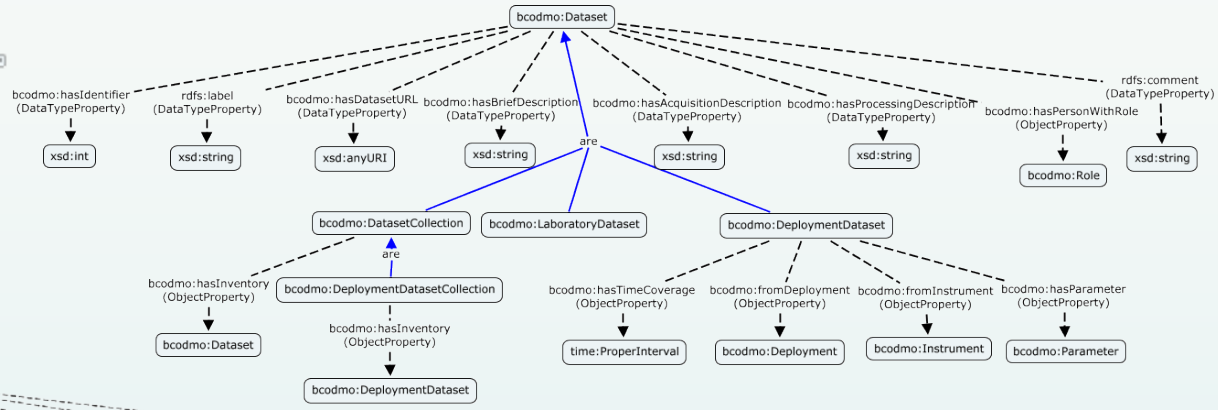
Ontology development

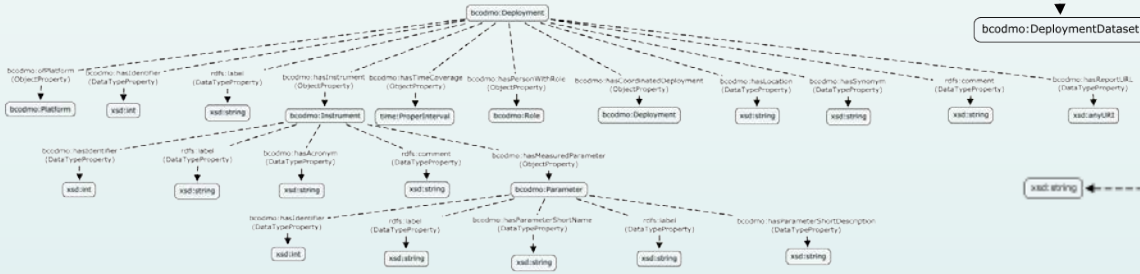
- organize all metadata into a single relational database (MySQL)
- identify, cleanup and map controlled vocabularies
- develop use cases based on database use examples
- do schema mapping (tools: Skype, Dimdim, CMapTools)
- populate the ontology with instances from the BCO-DMO MySQL database according to the schema mapping results (PHP, RDF)

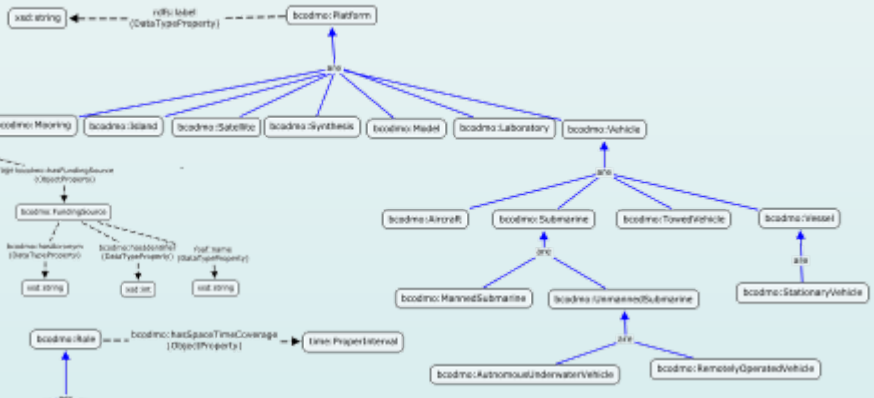
BCO-DMO Ontology

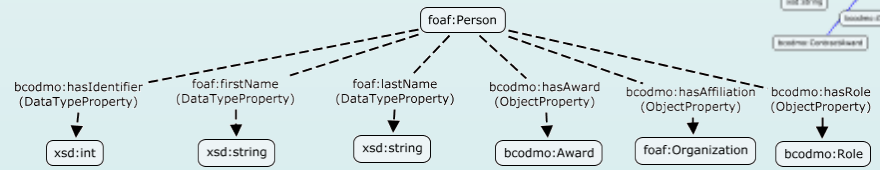


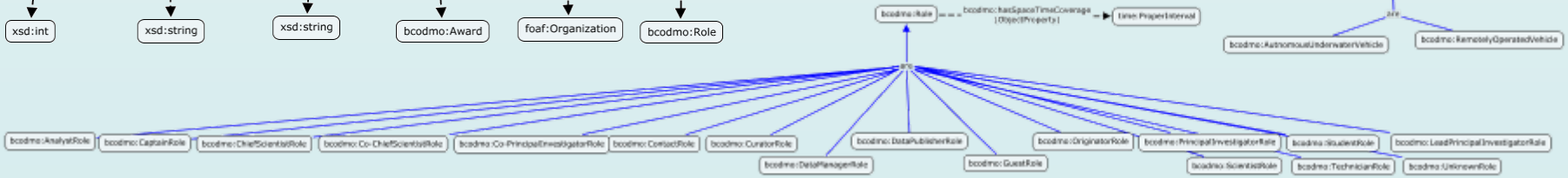




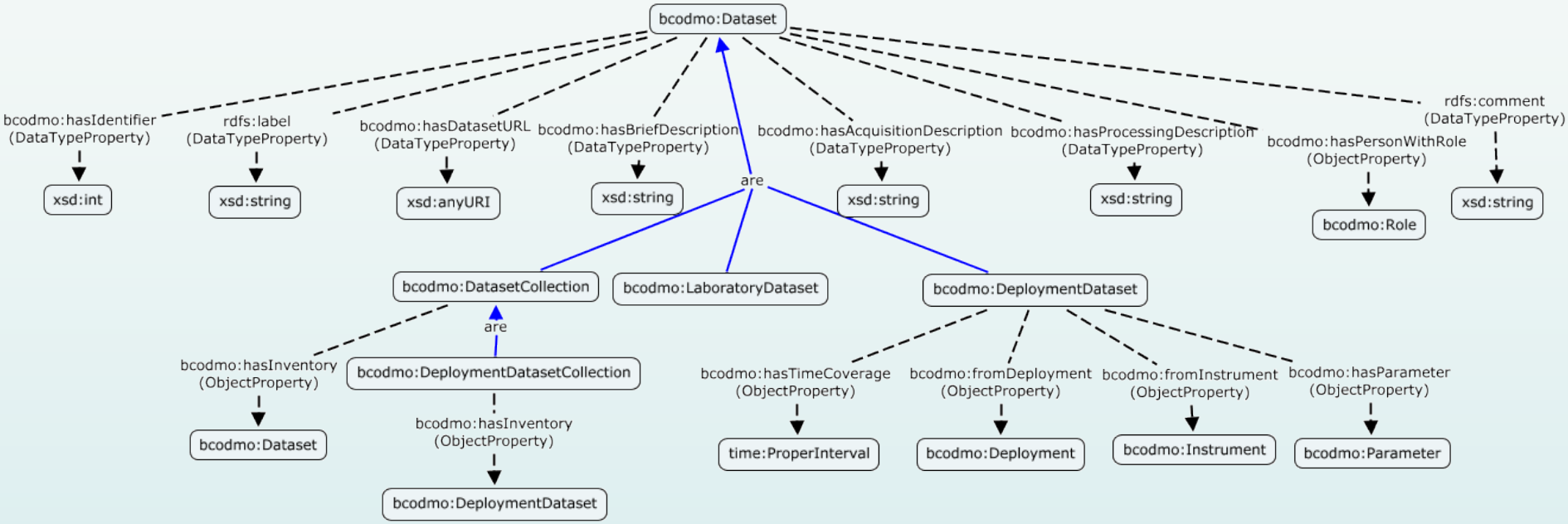








Dataset Concept Map from the BCO-DMO Ontology



What's Next? Vocabularies

- BCO-DMO Instrument Vocabulary
  complete the 'matching and mapping' of the remaining 'difficult' terms

- BCO-DMO Parameter Vocabulary
  cleanup terms to create controlled vocabulary, prepare for matching and mapping to SeaDataNet Parameter vocabularies

- Identify other BCO-DMO controlled vocabularies:
  geolocation (VLIZ Marine Gazetteer) terms
  units of measure (SI units as ISO 80000)

What's Next? Ontology

- Schema Mapping
  review use cases and complete the schema mapping
- Populating the Ontology
  with instances from BCO-DMO database
  (external RDF triple store)
- Harmonization of the BCO-DMO ontology
  check other ontologies for classes
  we can use
- Migrate "everything" to drupal 6


http://bco-dmo.org
thank you