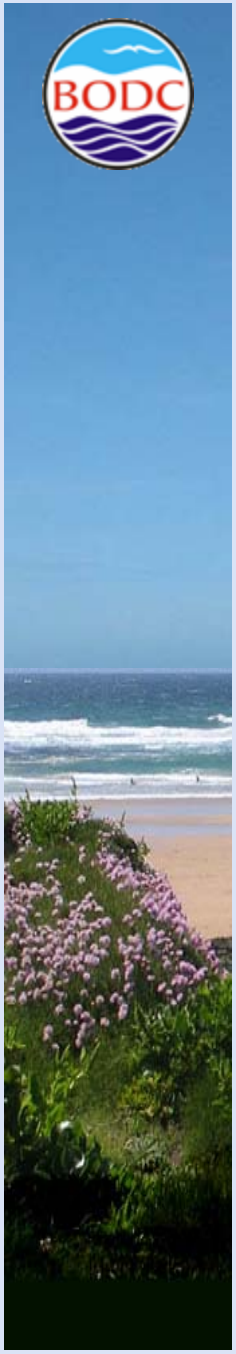
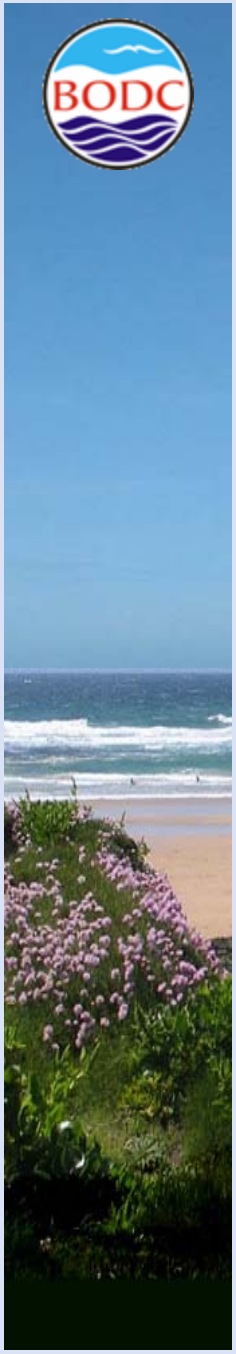


USNODC Seminar, September 2008

The Role of the Semantic Web in Oceanographic Data Management

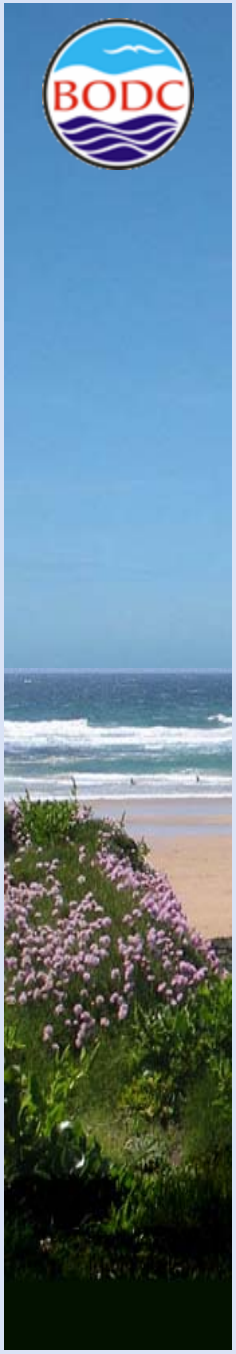
Roy Lowry

British Oceanographic Data Centre



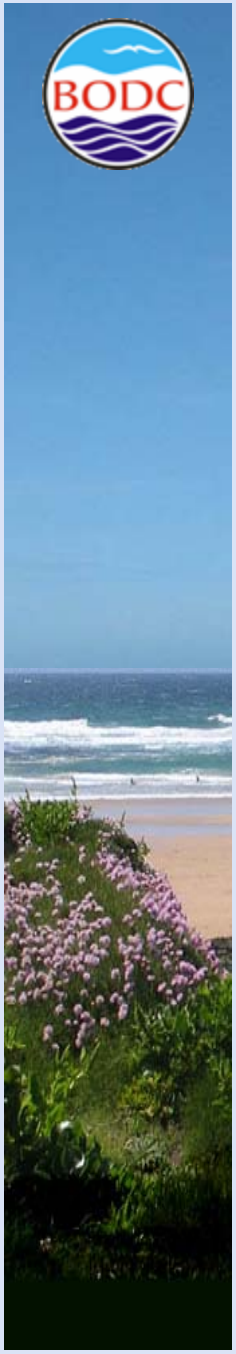
Presentation Overview

- The Semantic Web
- From codes to ontology
- The NERC DataGrid Vocabulary Server
- Technology Usage Examples
  - Semantic cross-walk
  - SeaDataNet metadata content verification



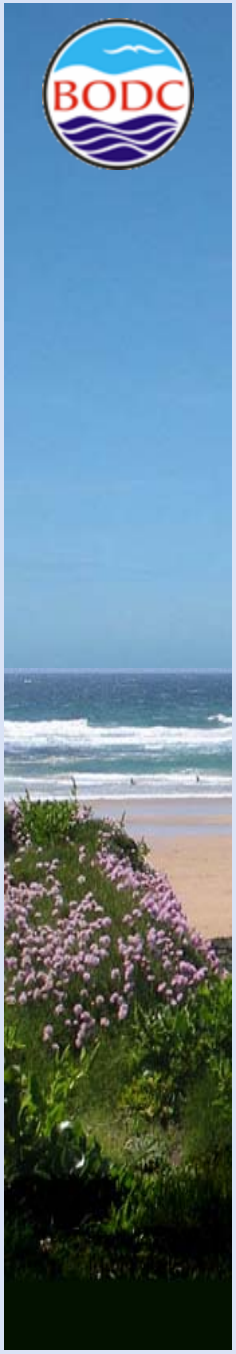
Semantic Web

- The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. (Tim Berners-Lee)
- In other words Web technology that delivers meaning for data and metadata
- A unit of data may be linked through a URN tag to a URL that returns an XML document stating what that data means and how it relates to other data



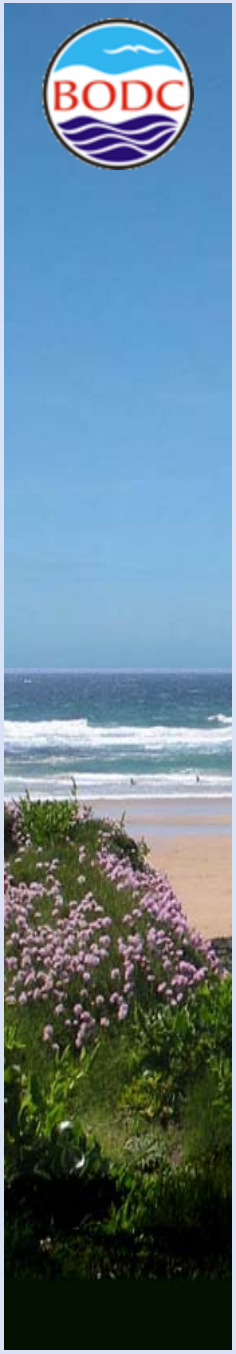
Semantic Web

- What does this mean in terms of oceanographic data management?
- We give data meaning in oceanographic management by tagging the data with metadata, often through the addition of codes to data streams
- Trouble is we don't always understand what each others' tags mean or how they relate
- The Semantic Web provides the framework for a distributed network delivering the information to overcome these misunderstandings



Codes to Ontology

- **What is a code?**
  - **My definition is:**
    - A tag attached to a data value that represents an object or information concept in the real world
- **Ideally**
  - **Objects and concepts are:**
    - Unambiguously defined
    - Universally understood
    - Unchanging

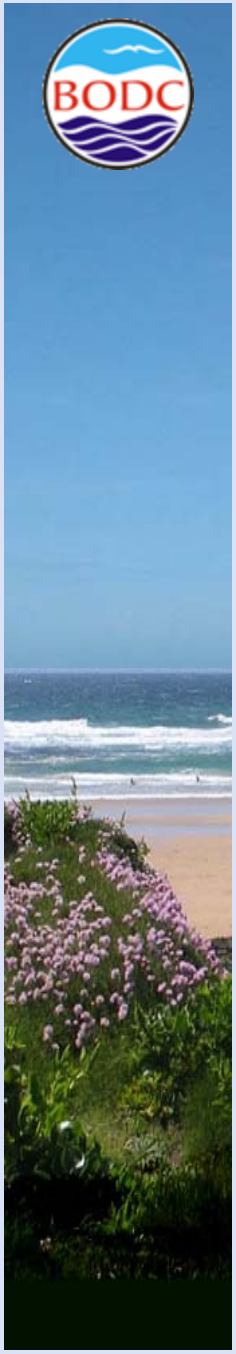


Codes to Ontology

- **Oceanographic data management is a long way from ideal**
  - **Long established practice has been to simply link a code to a short phrase of plaintext, such as**

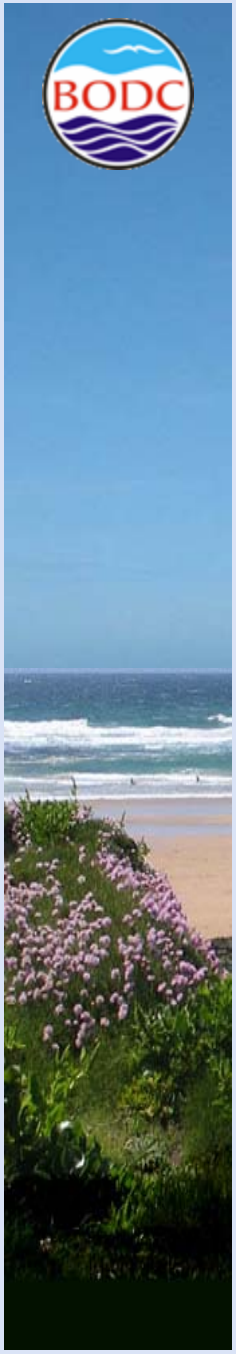
    * 74DS Discovery
    * 06 Germany

  - **The mapping between this plaintext and the real world has been somewhat flexible to say the least**



Codes to Ontology

- Let us consider ‘Discovery’ as an example
- What did a 1980s oceanographic data manager understand by Discovery?
- The British research vessel of course
- But which one?
  - Scott’s Antarctic Expedition ship
  - The NIO research vessel (Discovery II – it was written on her bow)
  - The IOS (now NERC) research vessel



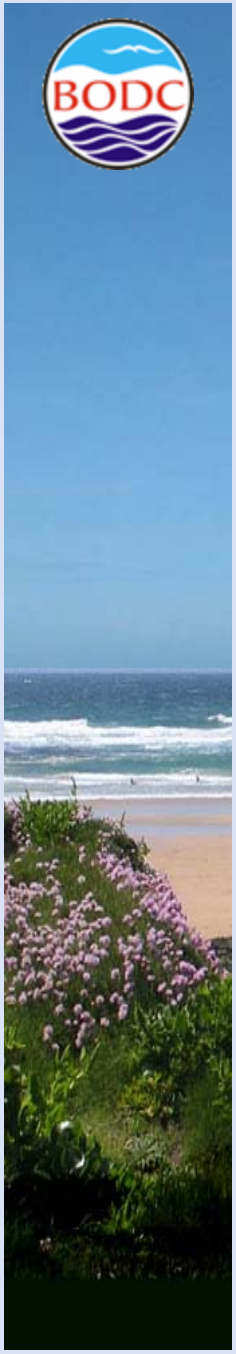
Codes to Ontology

- **ICES (i.e. Harry) said**
  - **74DI Scott's ship**
  - **74DS NIO and IOS/NERC ships**
- **NODC said**
  - **74DI Scott's ship**
  - **74DS NIO ship**
  - **74E3 IOS/NERC ship**
- **ICES now recognise 74E3 as part of platforms group rationalisation**
- **But are we out of the woods?…..**

Codes to Ontology

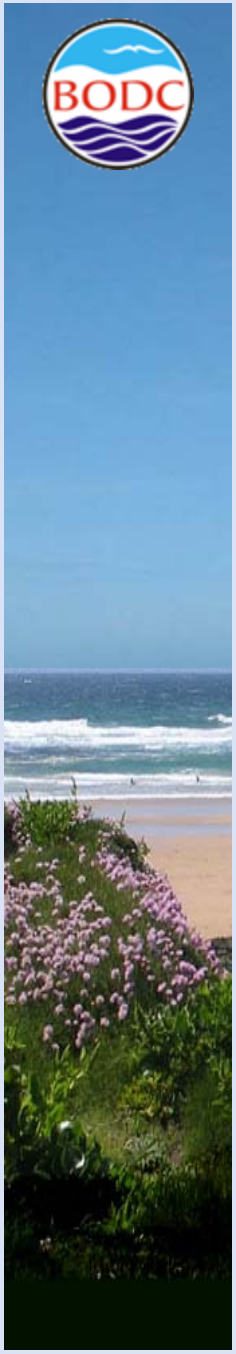
Are these the same ship?

Lloyds say they are, so according to recognised domain governance we're OK



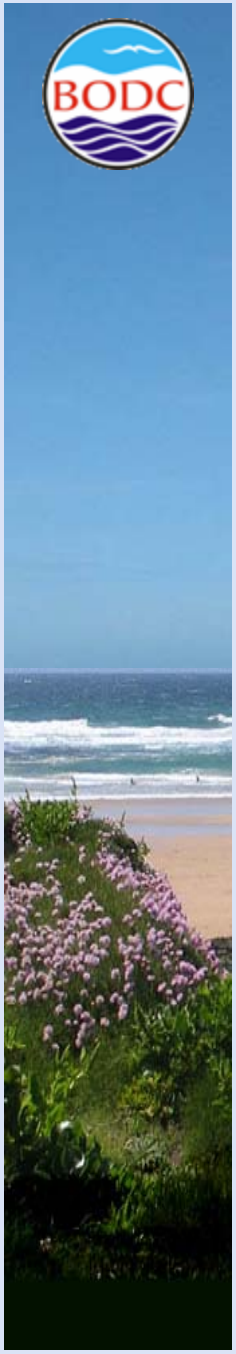
Codes to Ontology

- The problem is that our modelling of the real world has been grossly oversimplified
- Let us consider how we could model ships by metadata – i.e. develop a ‘ship’ class
- The fundamental physical entity is the ‘hull’ identified, except for small boats, by an IMO number



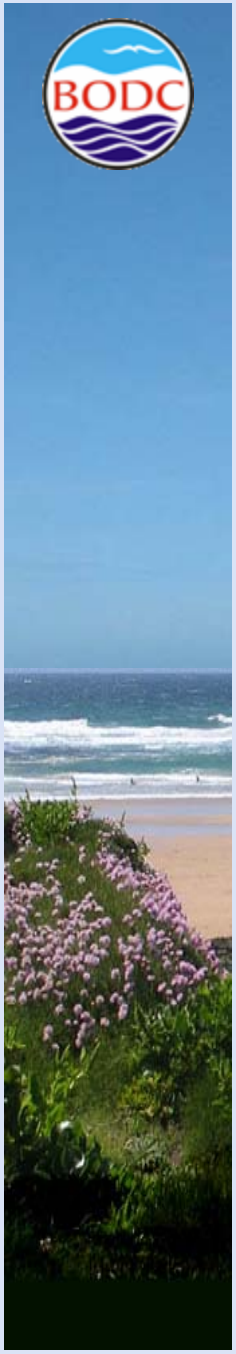
Codes to Ontology

- We could take IMO as the instance identifier (primary key in the relational world) of a class with the following attributes:
  - Name
  - Callsign and MMI number
  - Ownership and Flag
  - Vessel type classification
  - Size
  - Tonnage
  - Berths
  - Instrumentation configuration
- All of these can vary with time



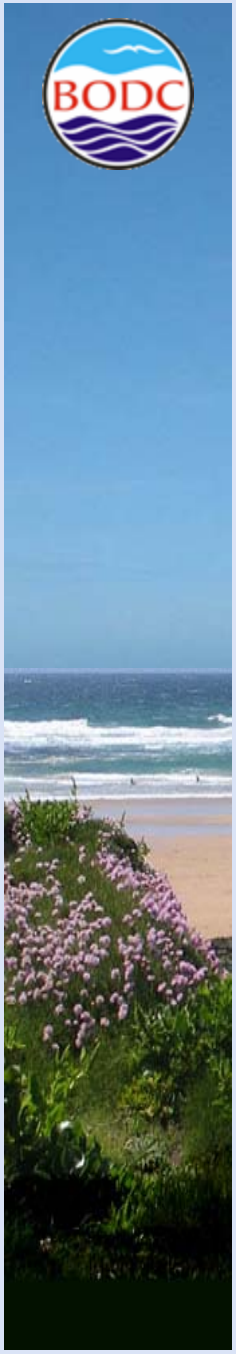
Codes to Ontology

- So, our metadata model needs to formalise multiple attribute sets, each labelled with a valid time window
- Accessing this information resource (ship domain ontology) requires an intelligent, time-aware interface (AI mediator)
- Or we can ‘cheat’ by redefining the entity as an instance of a set of attributes (hull, name, call sign, governance) and giving this a ‘ship code’
- Not ideal, but it’s legacy compatible and seems to work



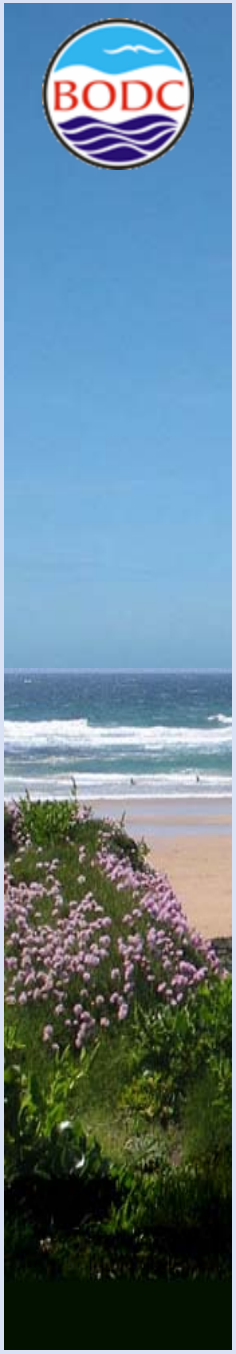
Codes to Ontology

- In BODC we are currently building an organisation ontology modelling name changes, mergers and dissolutions
- Fronted by functions implemented as an SQL extension
  - Current name: nmnow (code)
  - Previous name: nmthen (code,date)
  - History: nmall (code)



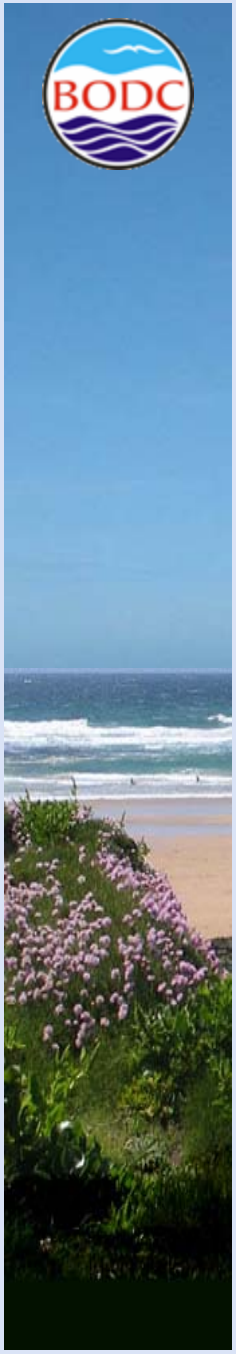
Codes to Ontology

- Building metadata knowledge resources requires large amounts of careful manual work
- Duplication of such work is a criminal waste
- Co-operation and sharing is the way to go
- The Semantic Web provides the infrastructure to make this possible



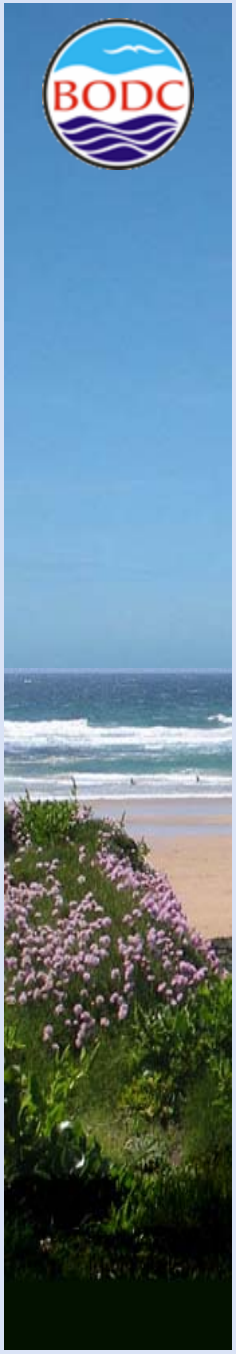
The NDG Vocabulary Server

- **This is a Semantic Web resource operated by BODC**
  - **Developed as part of the NERC DataGrid project**
  - **Adopted as the semantic element of the European Union SeaDataNet distributed data system**



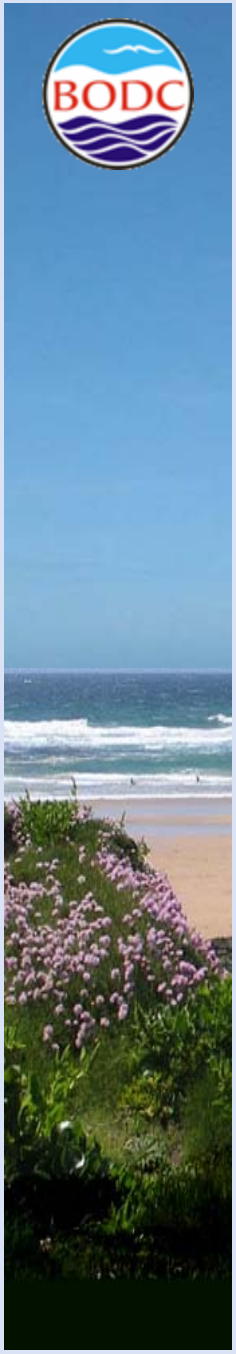
The NDG Vocabulary Server

- The server ‘payload’ is an XML document covering concepts or groups of concepts (optionally organised into vocabularies)
- Documents contain
  - Concept labels (names, abbreviations, URNs)
  - Concept definitions
  - Concept relationships to other concepts



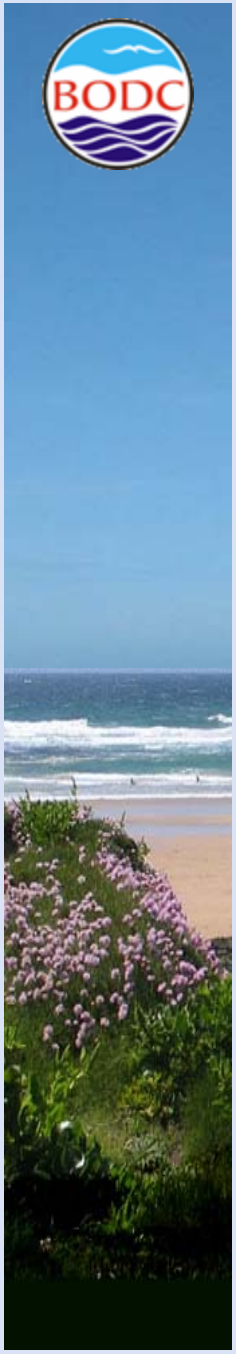
The NDG Vocabulary Server

- **Concepts are represented by URNs that have the form:**
  - **SDN:list_id:list_version:term_id, e.g.**
    - SDN:P021:23:PHYC
    - SDN:P021::PHYC (for current version)
  - **URNs resolve to URLs by simple string substitution**
    - SDN = http://vocab.ndg.nerc.ac.uk/term
    - P021 = P021 Null = current PHYC = PHYC
    - Giving http://vocab.ndg.nerc.ac.uk/term/P021/current/PHYC
  - **This returns the following XML document**



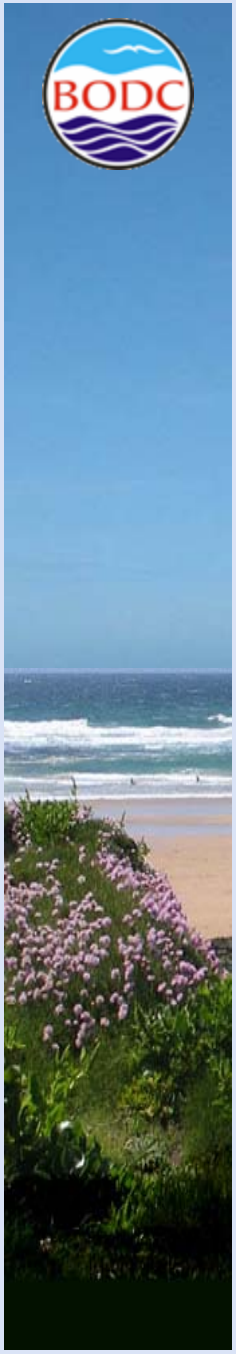
The NDG Vocabulary Server


```
<?xml version="1.0" ?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
- <skos:Concept rdf:about="http://vocab.ndg.nerc.ac.uk/term/P021/25/PHYC">
  <skos:externalID>SDN:P021:25:PHYC</skos:externalID>
  <skos:prefLabel>Phycobolin pigment concentrations in the water
    column</skos:prefLabel>
  <skos:altLabel>WC_PhycobolPig</skos:altLabel>
  <skos:definition>Concentration of phycobolin group pigments such as phycocyanin
    and phycoerythrin in the water column</skos:definition>
  <dc:date>2008-03-11T11:56:27.531+0000</dc:date>
  <skos:minorMatch rdf:resource="http://vocab.ndg.nerc.ac.uk/term/P041/4/G905" />
  <skos:broadMatch rdf:resource="http://vocab.ndg.nerc.ac.uk/term/P031/8/B035" />
  <skos:broadMatch rdf:resource="http://vocab.ndg.nerc.ac.uk/term/P041/4/G378" />
  <skos:broadMatch rdf:resource="http://vocab.ndg.nerc.ac.uk/term/P051/0/002" />
  <skos:broadMatch rdf:resource="http://vocab.ndg.nerc.ac.uk/term/P051/0/014" />
  <skos:narrowMatch
    rdf:resource="http://vocab.ndg.nerc.ac.uk/term/P011/79/PHYCSPP4" />
  </skos:Concept>
  </rdf:RDF>
```



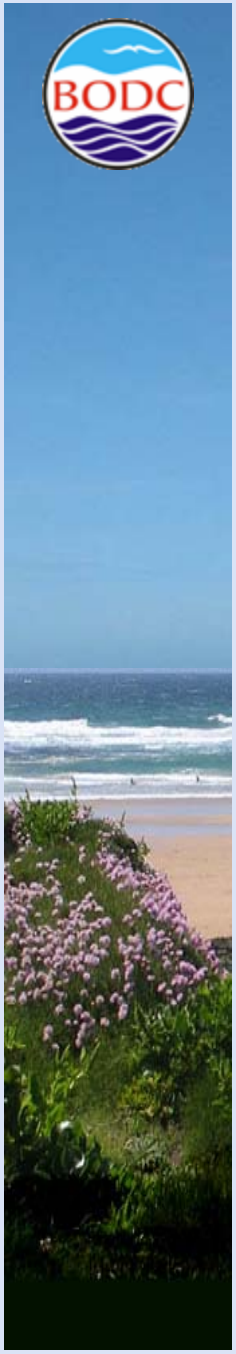
The NDG Vocabulary Server

- More sophisticated access is also possible through:
  - HTTP-POX web service calls
  - SOAP web service calls
  - Interface clients



The NDG Vocabulary Server

- **HTTP-POX service calls**
  - **Any API method may be invoked using an HTTP get call**
  - **Lists and terms specified in the get call parameters as URLs**
  - **Delivers an appropriate XML document (BODC-designed schema)**
  - **Documentation at** http://www.bodc.ac.uk/products/web_services/vocab/methods.html
- **SOAP web service calls**
  - **WSDL may be found at http://vocab.ndg.nerc.ac.uk/**
  - **Same output and documentation as HTTP-POX**



The NDG Vocabulary Server

➢ Interface clients

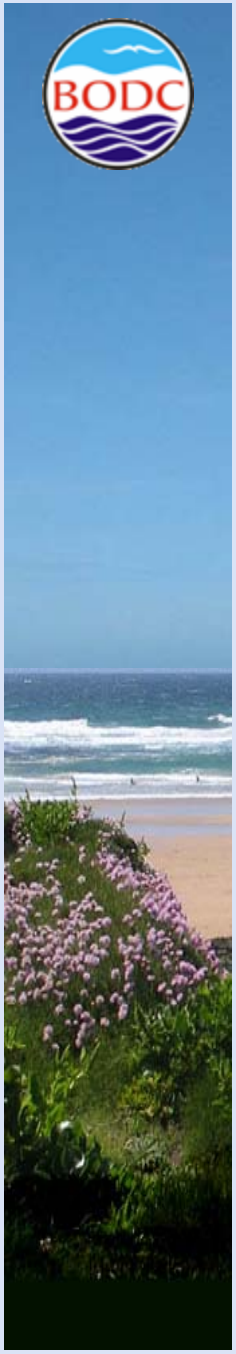
- **Maris client set up for SeaDataNet at http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx**

- **BODC clients at http://vocab.ndg.nerc.ac.uk/ cover more vocabularies if interests extend beyond SeaDataNet**

The NDG Vocabulary Server

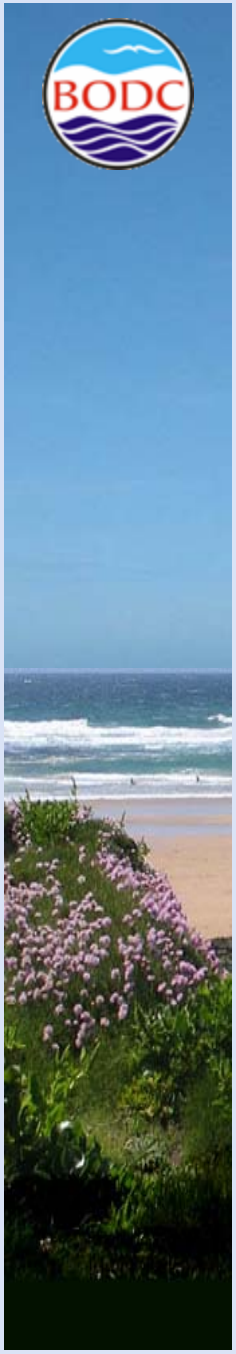
- **Maris client home page**

List key	Long name	Short name	List version	List last modified
C161	International Hydrographic Bureau (1953) sea areas	IHB Sea Areas	0	1/1/1954 12:00:00 AM
	Terms used for sea areas from International Hydrographic Bureau, Limits of Oceans and Seas (Special Publication No. 23), 3rd edition 1953.			
C162	IOC cruise summary report sea area extensions	ROSCOP sea area extensions	0	1/1/1982 12:00:00 AM
	Terms for sea areas added to the IHB (1953) list to provide the standardised sea area descriptions used in IOC cruise summary form reporting. Celtic Sea was formally adopted by IHB.			
C16	SeaDataNet Sea Areas	SeaDataNet Sea Areas	0	3/1/2007 12:00:00 AM
	SeaDataNet Sea Areas			
C173	Partnership for Observation of the Global Ocean ships of interest	POGO ships	11	1/12/2008 12:00:00 AM
	Research vessels deemed to be of interest to POGO. 'Of interest' is defined as active ocean-going research vessels greater than 60m in length			
C174	SeaDataNet Cruise Summary Report ship metadata	SDN CSR ships	1	2/21/2008 12:00:00 AM
	Ship instances (a hull operating under a given name and governance type) used in SeaDataNet CSR forms including metadata in the definition to allow reliable mapping to ship hull databases.			
C17	PLATFORM CODE	PLATFORM CODE	37	2/21/2008 12:00:00 AM
	PLATFORM CODE			
C320	International Standards Organisation countries	ISO countries	1	8/21/2007 11:00:00 PM
	ISO country codes from ISO3166-1 list taken from www.iso.org on 22/08/2007.			
C321	International Standards Organisation deprecated country codes	ISO deprecated countries	1	8/21/2007 11:00:00 PM
	Deprecated ISO country codes from the ISO3166-3 list.			
C32	ISO country codes from ISO3166 list	ISO country codes from ISO3166 list	1	8/21/2007 11:00:00 PM
	ISO country codes from ISO3166 list			
C342	Monitoring activity rationale	Monitoring rationale	2	12/6/2007 12:00:00 AM
	Terms describing the reasons why a monitoring activity was undertaken.			
C361	Monitoring activity legislative drivers	Monitoring drivers	4	2/15/2008 12:00:00 AM
	Legislative acts, agreements and treaties that have provided the impetus for monitoring activities to be undertaken.			
C371	Ten-degree Marsden Squares	Marsden-10	1	8/2/2007 11:00:00 PM
	Labels applied to areas of ten degrees latitude by ten degrees longitude in the Marsden Square system.			
C372	Five-degree Marsden Squares	Marsden-5	1	8/2/2007 11:00:00 PM
	Labels applied to areas of five degrees latitude by five degrees longitude in the Marsden Square system.			
C373	One-degree Marsden Squares	Marsden-1	1	8/8/2007 11:00:00 PM
	Labels applied to areas of one degree latitude by one degree longitude in the Marsden Square system.			
C37	Marsden Squares	Marsden Squares	3	8/8/2007 11:00:00 PM
	Marsden Squares			
C381	Ports Gazetteer	Ports Gazetteer	7	11/14/2007 12:00:00 AM
	Geographic locations from which a cruise may begin or end			
C382	Ports Gazetteer Deprecated Entries	Deprecated Ports	4	11/14/2007 12:00:00 AM
	Entries that have been defined in the C381 cruise start and end point gazetteer, but have been deprecated			



The NDG Vocabulary Server

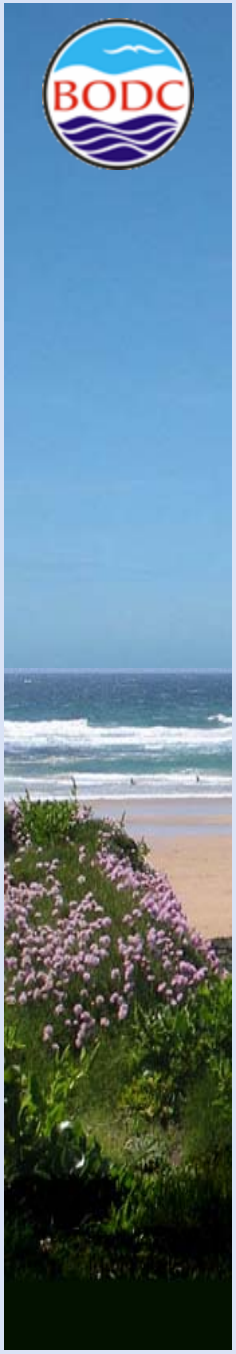
- **Server Contents (2008-08-21)**
  - **112 public lists**
  - **122603 concepts**
  - **78123 mappings (RDF triples)**
- **Server Usage 2008 (to 2008-08-21)**
  - **2233803 total hits (2000000 of these attributable to robots)**
  - **37462 vocabulary catalogue hits**
  - **50458 vocabulary list downloads**
  - **2085 vocabulary mapping queries**



The NDG Vocabulary Server

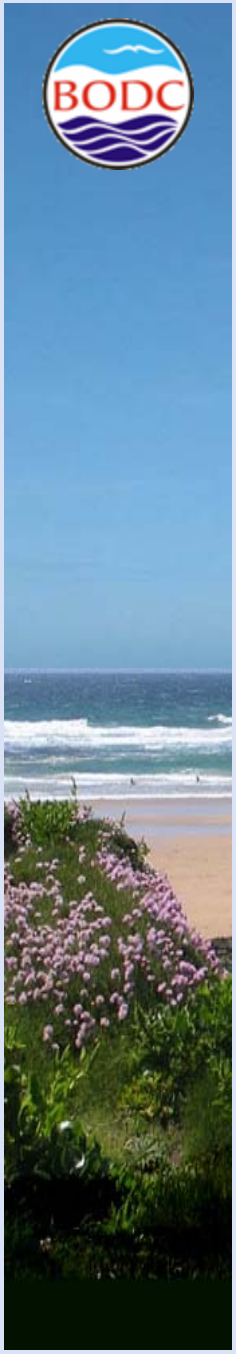
➢What's Wrong With It?

- **Historic version serving not implemented**
  - Current version served whatever version is requested
- **Predicates (based on SKOS mappings) semantically limited**
  - More suited to a thesaurus than an ontology
  - A richer predicate set exists in the triple store, but cannot be served without WSDL changes



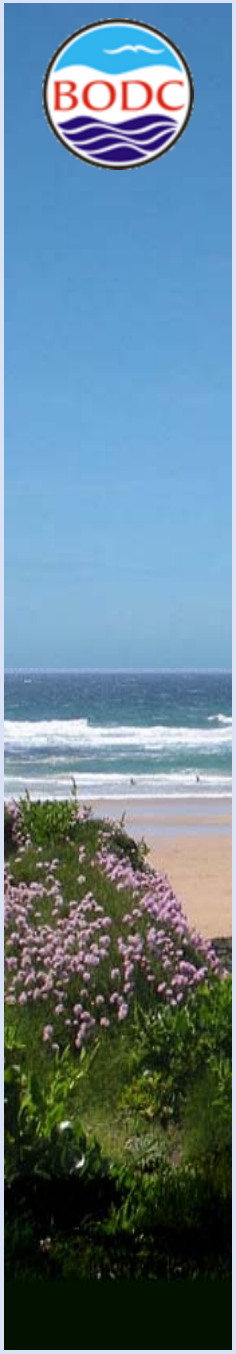
The NDG Vocabulary Server

- **What's Wrong With It?**
  - **Vocabularies not labelled with content governance authority**
  - **Mappings restricted to concepts within the server**
    - If a vocabulary is to be included in a mapping then it must be loaded in the server
- **Development continues to address these issues**



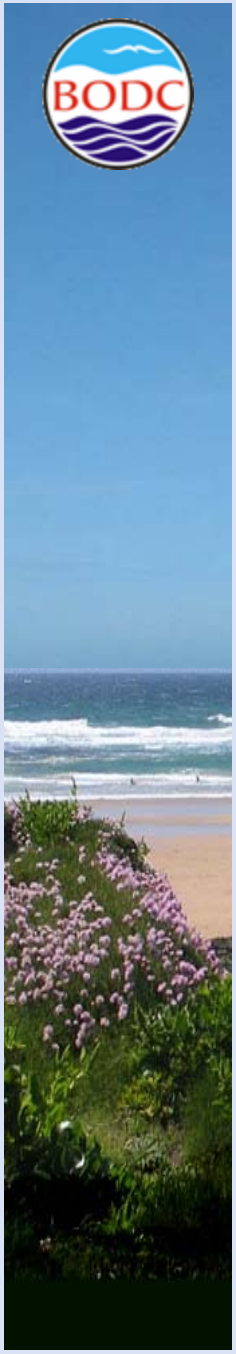
Semantic Crosswalk Use Case

- BODC wishes to produce a GCMD DIF document from an EDMED V1.2 document
- The “parameter” sections of the two documents are populated using different vocabularies (BODC PDV and GCMD Science Keywords)
- This situation was usually addressed by having no parameter section in the output document
- We can now do better…..



Semantic Crosswalk Use Case

- **A list of BODC PDV terms as parameter URNs is obtained from the EDMED document, for example:**
  - SDN:P021:24:TEMP, SDN:P021:24:PSAL , SDN:P021:24:CPWC
- **This may then translated into a list of URLs**
  - http://vocab.ndg.nerc.ac.uk/term/24/TEMP
  - http://vocab.ndg.nerc.ac.uk/term/24/PSAL
  - http://vocab.ndg.nerc.ac.uk/term/24/CPWC

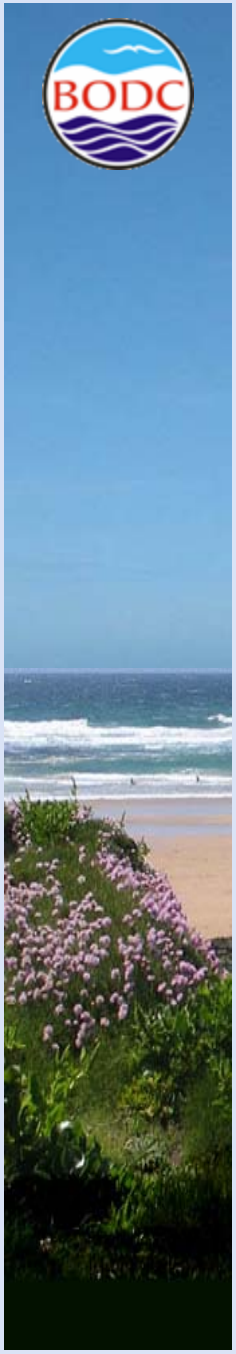


Semantic Crosswalk Use Case

This list may be rolled into an HTTP get request thus:

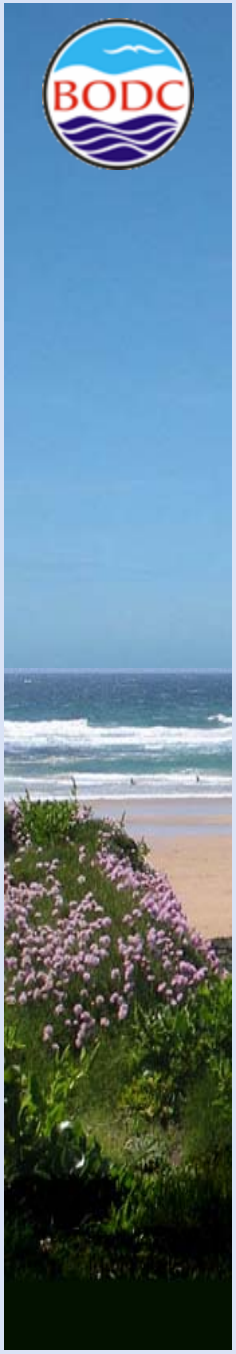
- **http://vocab.ndg.nerc.ac.uk/axis2/services/vocab/getRelatedRecordByTerm?subjectTerm=http://vocab.ndg.nerc.ac.uk/term/P021/current/TEMP&subjectTerm=http://vocab.ndg.nerc.ac.uk/term/P021/current/PSAL&subjectTerm=http://vocab.ndg.nerc.ac.uk/term/P021/current/CPWC&objectList=http://vocab.ndg.nerc.ac.uk/list/P041/current&predicate=255&inferences=true**

- **An XML document is returned containing the GCMD Science Keywords that map to the three BODC terms as both text strings and URLs**
- **The document may be reformatted using XSLT or XQuery to generate the "parameters" section for the DIF**



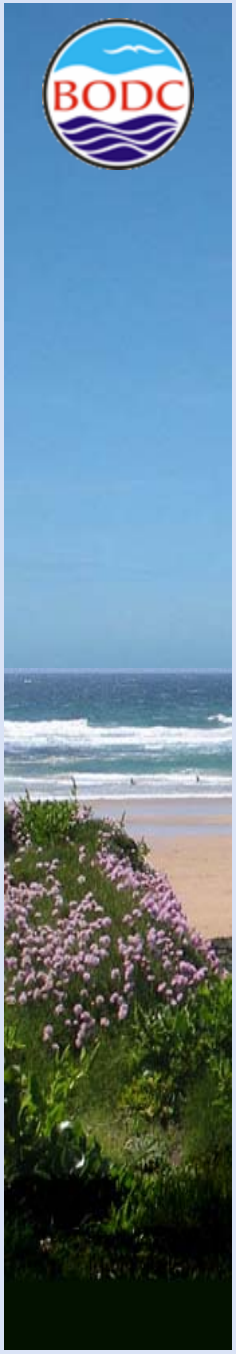
Metadata Content Verification

- During SeaSearch an EDMED submission was repeatedly rejected by BODC, but the originators insisted there was nothing wrong with it
- The originators had built the document using vocabularies that had developed locally because no workable central governance was in place
- In SeaDataNet we were determined to prevent a recurrence of this situation by:
  - Installing vocabulary governance that works
  - Providing tools for partners to verify metadata CONTENT at source against master vocabularies



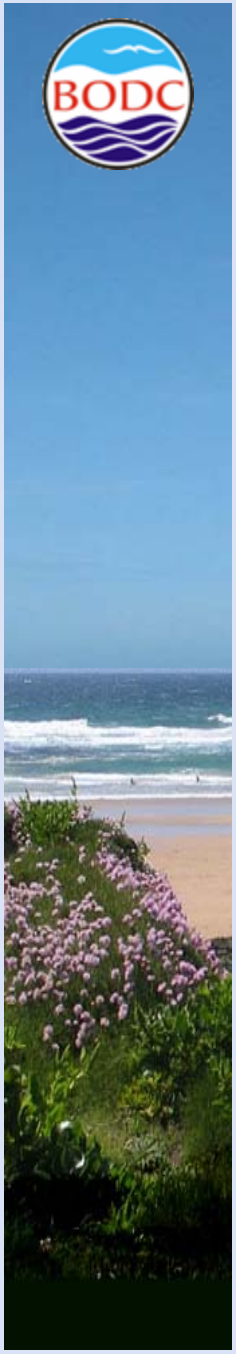
Metadata Content Verification

- **The content verification uses Semantic Web technology:**
  - **The SeaDataNet XML metadata schemas comprise two parts:**
    - The base schema describing the document structure
    - Schema extension coded in Schematron describing controlled field content



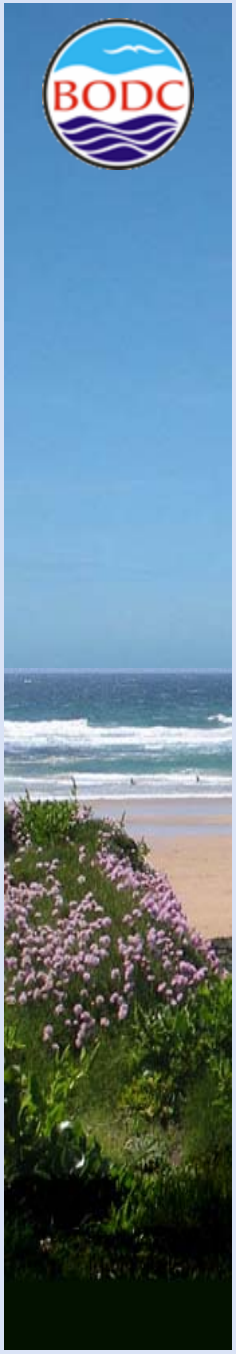
Metadata Content Verification

- The base schema is served using a conventional online change control management system (BSCW)
- The schema extensions are added by a Web Service operated by the Russian NODC
- This builds the Schematron code using documents generated by Vocabulary Server calls based on URNs encoded in the metadata



Metadata Content Verification

- **SeaDataNet partners may either:**
  - **Download the extended schema and verify their XML documents using generic XML tools like Oxygen**
  - **Upload their documents to and verify against the extended schema using a tool provided by the Russians.**
- **This significantly accelerates ingestion because issues of both structure and content are resolved prior to submission**



That’s All Folks

- Thank you for your attention
- Questions?