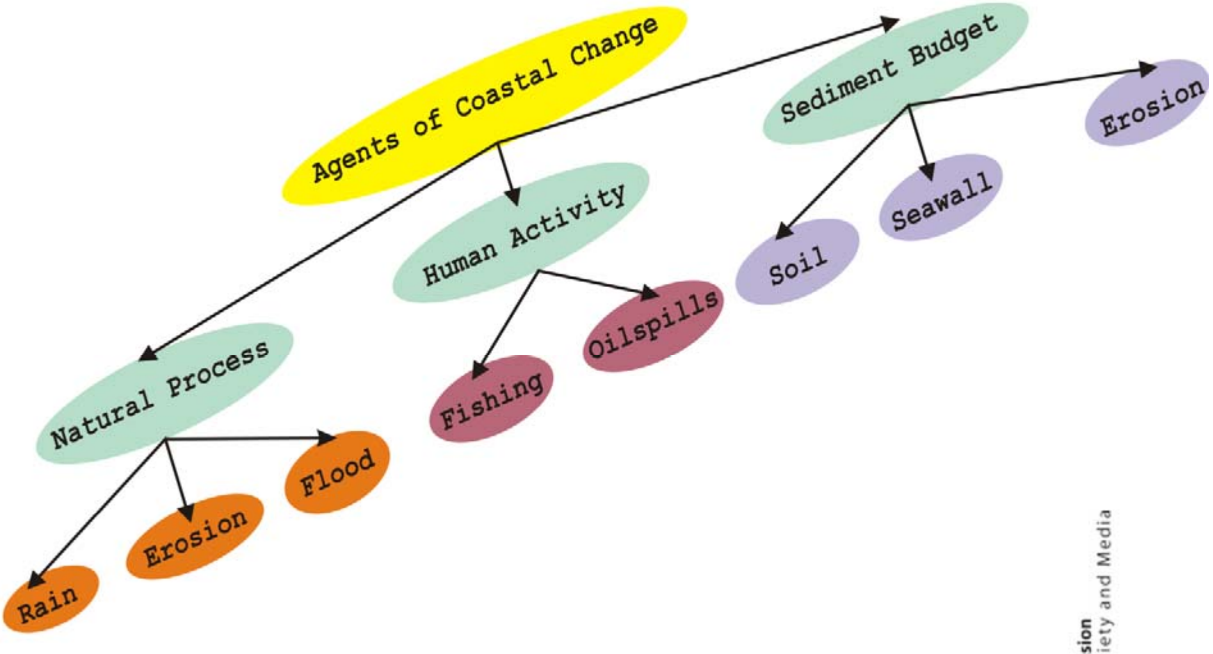
International Coastal Atlas Network Cookbook: Understanding Semantics



European Commission
Information Society and Media

Concept map taken from the ICAN Coastal Erosion Thesaurus - http://vocab.nerc.ac.uk/scheme/ICANCOERO/current/

Table of Contents

Introduction

"If HTML and the [World Wide] Web made all the online documents look like one huge book, [semantics] will make all the data in the world look like one huge database"

Tim Berners-Lee[1]

If data in a distributed system are to be understood elsewhere in that system, or externally to the system, they must be labelled (or “marked up”) using a common set of meaningful terms or phrases. These common phrases must be consistent throughout the full data system, or there must be a means of translating between the phrases used at different points of the system, using common “semantics”. Semantics is the study of meaning; it focuses on the relationships between words and what they stand for or mean. The aim of the “semantic web” is to provide these consistent phrases and to define the relationships in a formal manner, resulting in what is often called a “knowledge organization system”.

This document provides a tutorial for those who wish to investigate and make use of these technologies, aimed specifically at members of the International Coastal Atlas Network community and more generally at environmental scientists and data managers.

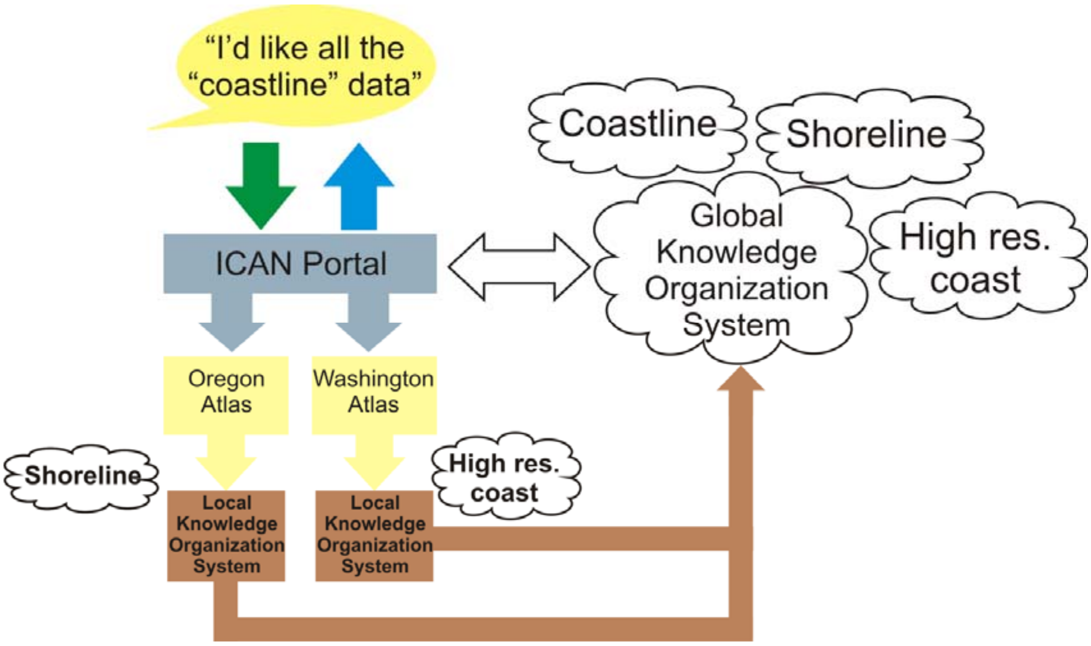
Why use a “knowledge organization system”?

One scenario for using knowledge organization systems in the International Coastal Atlas Network[2] (ICAN) is to search through the local atlases for a given data keyword from a central portal. For example, as illustrated below, a user arrives at the ICAN portal and request “coastline” data. The portal software is connected to a global knowledge organization system which is aware that “coastline” is related to both “shoreline” and “high resolution coastline”. The user request and this information from the global knowledge organization system are then passed on to the local atlases which search on “coastline”, “shoreline” and “high resolution coastline”. The local atlases then return the relevant data to the portal and then to the user. This is an implementation of so-called “smart-search”[3].

[1] Berners-Lee, T. (1999) *Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor*. Orion Business. ISBN-100752820907

[2] http://ican.science.oregonstate.edu/

[3] Latham, S. E.; Cramer, R.; Grant, M.; Kershaw, P.; Lawrence, B. N.; Lowry, R.; Lowe, D.; O'Neill, K.; Miller, P.; Pascoe, S.; Pritchard, M.; Snaith, H.; Woolf, A. (2009) The NERC DataGrid services. *Philosophical Transactions of the Royal Society A*, 367 (1890). 1015-1019.

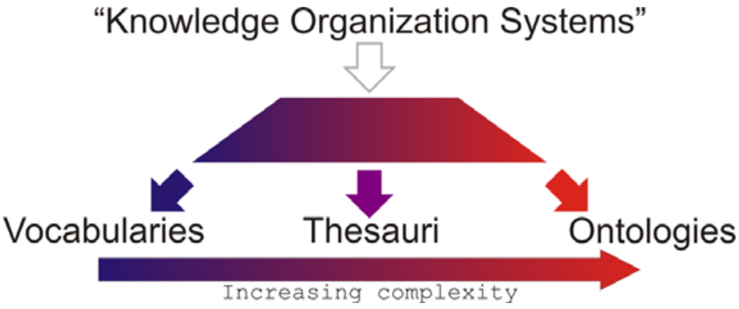

"I'd like all the "coastline" data"
Coastline
Shoreline
ICAN Portal
Global Knowledge Organization System
High res. coast
Oregon Atlas
Washington Atlas
Shoreline
High res. coast
Local Knowledge Organization System
Local Knowledge Organization System

A diagram illustrating one use for knowledge organization systems in the ICAN community.

Other uses of knowledge organization systems include populating metadata elements with standardized content which can be verified and validated by software services; dynamically populating drop down lists in websites and software applications; dynamically moving a metadata record from one metadata scheme to another; and the validation of input parameters and their associated units in Open Geospatial Consortium Web Processing Services.

What are vocabularies, thesauri and ontologies?

Knowledge organization systems fall broadly into three groups: vocabularies, thesauri and ontologies. These three groups show increasing complexity in their structure as illustrated in the diagram below.



The "semantic spectrum" shows the increasing complexity of different forms of knowledge organization system. After McGuinness (2003)[4].

A vocabulary can be either a list of terms or a list of terms and some text providing a definition of the term. A vocabulary ensures that terms are used, and spelt, consistently. A vocabulary can be extended in its power by providing definitions of concepts.

Thesauri expand the knowledge contained within a vocabulary by adding information about the relationships between the terms of the vocabulary. These relationships fall broadly into three categories:

- Synonyms – the current term is synonymous with a given, different term. e.g. "dogs" is synonymous with "canines".
- Broader relations – the current term has a more specific definition than a given different term. e.g. "dogs" has a broader relationship to "pets"
- Narrower relations – the current term has a less specific definition than a given different term. e.g. "dogs" has a narrower relationship to "terriers"

In a more complex thesaurus, the concepts at the top of the hierarchy of broader and narrower relations may be stated explicitly, rather than being inferred by software agents. A well known example of this form is the Yahoo! web directory[5] or the categorisation of auctions on the eBay homepage[6]. eBay has terms such as "Antiques", "Coins" and "Sporting Goods" as the top level in its hierarchy. Narrower terms sit below these, for example "Sporting Goods" contains "Football", "Golf" and "Sailing". These terms sit above those which are narrower still, "Sailing" having such narrower terms as "Clothing & Shoes", "Life Jackets" and "Rope". In the context of environmental sciences, the Global Change Master Directory[7] can be seen to work in this way. For example, "Oceans" is at the top level, with "Coastal Processes" beneath it and terms such as "Beaches" and "Coastal Elevation" beneath that.

These more complex thesauri also introduce a fourth category of relationship between concepts, that of a "loose relationship". That is where two terms have a relationship that is not of the broader or narrower type or a synonymous relationship, e.g. "domesticated dogs" are "loosely related" to "wild dogs". These loose relationships may allow different pathways to the

[4] Deborah L. McGuinness. (2003) Ontologies Come of Age. In Dieter Fensel, James Hendler, Henry Lieberman, and Wolfgang Wahlster (eds). *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. Massachusetts Institute of Technology Press.

[5] http://dir.yahoo.com/

[6] http://www.ebay.com/

[7] http://gcmd.nasa.gov/

discovery of a term, making the resource what is known as “orthogonal”. For example, eBay has “Walking, Hiking, Trail” in its “Fashion” auction categories and “Boots & Shoes” in its “Sporting Goods” auction categories. If these two were loosely mapped a search for “walking boots” could yield auction results from both categories.

A thesaurus may be expanded to an ontology by declaring a term to belong to a particular class; or the addition of property information to the term; or the restriction of values that data associated with the term may take. An ontology class is used to define a type which can be used to group related terms. For example, if eBay defined the class of “auction” particular individual terms belonging to the “auction” class could be “English auction”, “blind auction” or “Dutch auction”.

How to discover existing knowledge organization systems?

Can I reuse existing resources?

Where possible it is best to make use of existing knowledge organization systems. This increases the ability to reuse data across systems, known as interoperability. If the reuse of existing systems is not an option, the section below explains how to generate a new knowledge organization system. Any new system should have some specified relationships to an existing system to promote interoperability and flexibility (see page 13). Details of how to access an existing knowledge organization system relevant to the International Costal Atlas Network are provided on page 12 of this document.

It is also possible to extended existing resources by creating mappings between them and other resources. This activity is described on page 13, below.

Where might I find existing knowledge organization systems?

In order to reuse existing resources, it is essential to know where to find them and how to asses their quality. Existing resources which may be of interest can often be found in ontology registries, for instance the Marine Metadata Interoperability Ontology Registry and Repository[8] or the NERC Vocabulary Server[9]. The former has a search facility on its home page; the latter may be searched most easily through the SeaDataNet vocabulary pages hosted by Maris[10]. Both of these systems provide publication mechanisms for knowledge organization systems which may be created by a range of authorities, and the creating authority is acknowledged in the systems’ output. An additional benefit of these systems is that they provide versioning of the content of the knowledge organization systems that they serve.

Other resources that are of interest to the Earth Sciences domain exist outside of these registry systems. These include the NASA’s Global Change Master Directory[7] and Semantic Web for Earth and Environmental Terminology[11]; the European Environment Agency’s General

[8] http://mmisw.org/orr/
[9] http://vocab.nerc.ac.uk/
[10] http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx/
[11] http://sweet.jpl.nasa.gov/ontology/

Multilingual Environmental Thesaurus (GEMET)[12]; the GeoSciML vocabularies[13]; and the United States Geological Survey thesaurus[14].

When considering the use of an existing knowledge organization system, the key things to look for are: an individual web address (or URL) to each term defined – this is how you will mark up your metadata; a well documented version control system; and an authoritative body in control of the content of the KOS.

How to define the content of a knowledge organization system?

What is the scope of the knowledge organization system?

While it might be tempting to want to describe and define every imaginable concept in a new knowledge organization system, this would be a very time consuming and frustrating process, and would not make best use of other, pre-existing resources. Instead, it is much better to take the time to identify the specific domain that needs to be described by the terms you wish to define, for example coastal erosion, or names and extents of beaches. In this way work in building the knowledge organization system is tightly defined and the content is coherent, well understood and should not replicate existing resources.

Identifying the content

How narrow or broad should a term definition be?

The challenge of integrating data and information of different kinds at different levels of detail is well defined in computer science literature[15,16]. In the area of semantics on the World Wide Web, the level of detail a term can describe is known as its granularity. For a given level of a knowledge organization system the definitions of a term may be as broad or as narrow as is necessary, as long as they are not ambiguous.

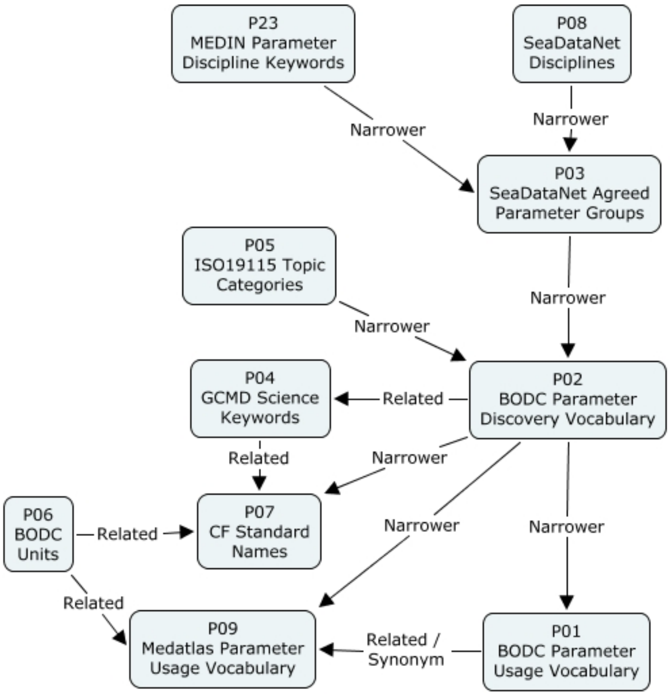
However, when building a hierarchical thesaurus, it is important that concepts defined at the same level of the hierarchy maintain a similar degree of granularity. If the thesaurus is imagined as a pyramid, making a concept at a given level too narrow or broad in its definition is like placing a too small or too large brick in the wall of the pyramid, and makes the structure unstable. For example, “body of water” should not sit at the same level as “lake” or “reservoir”, as these are terms with a narrower relationship or a finer granularity.

Linking term definitions together

As described above, the definition of terms by themselves is useful but the impact of the work can be greatly extended by providing relationships which link the terms together to form networks of knowledge. This enhances the ability of a user to find data labelled with a given

[12] http://www.eionet.europa.eu/gemet/
[13] http://srvgeosciml.brgm.fr/eXist2010/brgm/client.html
[14] http://www.usgs.gov/science/about/
[15] Fonseca, F., Egenhofer, M., Davis, C., and Câmara, G. (2002) Semantic Granularity in Ontology-Driven Geographic Information Systems. *AMAI Annals of Mathematics and Artificial Intelligence - Special Issue on Spatial and Temporal Granularity* 36(1-2): 121-151.
[16] Yan, X., Lau, R.Y.K, Song, D., Li, X., Ma, J. (2011) Towards a Semantic Granularity Model for Domain Specific Information Retrieval. *ACM Transactions on Information Systems (TOIS)*. In press.

term or to translate the metadata from one mark up scheme to another. Relationships can be thought of simply as broader and narrower (for example, in the diagram below the BODC Parameter Discovery Vocabulary is narrower than the SeaDataNet Agreed Parameter Groups and vice versa); loosely related (the BODC Parameter Usage and MEDATLAS Parameter Usage vocabularies are of similar granularity and are linked this way); and synonyms where two terms may be used interchangeably.


P23 MEDIN Parameter Discipline Keywords
P08 SeaDataNet Disciplines
Narrower
Narrower
P03 SeaDataNet Agreed Parameter Groups
P05 ISO19115 Topic Categories
Narrower
Narrower
P04 GCMD Science Keywords
Related
P02 BODC Parameter Discovery Vocabulary
Related
Narrower
P06 BODC Units
Related
P07 CF Standard Names
Narrower
Narrower
Related
P09 Medatlas Parameter Usage Vocabulary
Related / Synonym
P01 BODC Parameter Usage Vocabulary

An example from the NERC Vocabulary Server[9] to show how identifying relationships between terms builds a network of parameter definitions.

Ensuring the quality of the content of the Knowledge Organization System

There are two aspects to providing quality assurance, or governance, for a knowledge organization system. The first is to ensure the quality of the content of the knowledge organization system. This includes the names and definitions of terms and the relationships between the terms. A well tested mechanism for managing content governance is setting up an e-mail list of interested parties on which requests for new terms and mappings can be discussed. This is the model which has been implemented by: the Climate and Forecast[17] netCDF metadata conventions group; the SeaDataNet and MarineXML Vocabulary Content Governance Group (SeaVoX)[18]; and the NETMAR ontology governance body[19]. The role of the content governance group is analogous to the International Organization for Standardization (ISO) definition of a "control body"[20].

[17] http://cf-pcmdi.llnl.gov/
[18] https://www.bodc.ac.uk/data/codes_and_formats/seavox/
[19] http://netmar.nersc.no/
[20] http://www.dgiwg.org/Terminology/faq-other.php

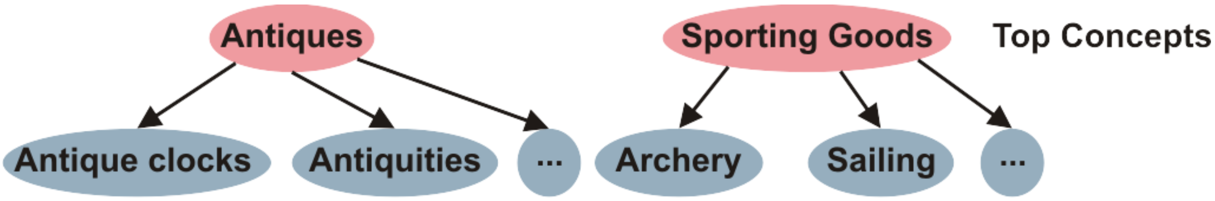
The second aspect is assuring the technical quality of the system. This includes ensuring that the knowledge organization system is available with the greatest possible up-time; the representation of the system is valid in the chosen scheme (e.g. extensible markup language, XML); and the various versions of the concepts, collections and scheme are maintained and accessible. For example, within the NETMAR project this technical governance is provided by the British Oceanographic Data Centre as the developer and maintainer of the NERC Vocabulary Server (NVS). The role of the technical governance group is analogous to the ISO definition of a "register manager"[20].

Making the content available

Simple Knowledge Organization System

The NETMAR project's knowledge organization systems are built upon the World Wide Web Consortium's Simple Knowledge Organization System[21] (SKOS) standard. SKOS is designed to provide a method for the online publication of controlled vocabularies and thesauri. NETMAR publishes two International Coastal Atlas Network thesauri and an Oregon Coastal Atlas thesaurus as XML documents using the SKOS standard. A brief overview of SKOS is therefore provided below.

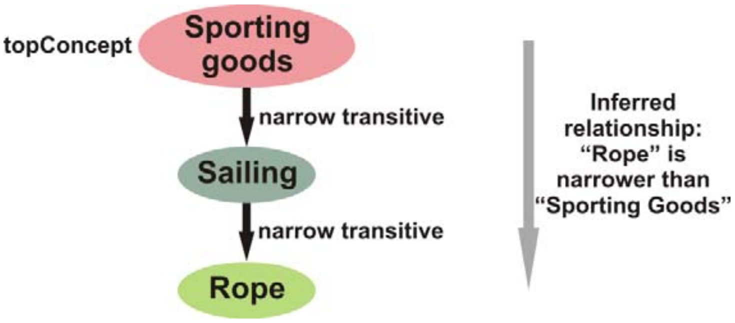
SKOS is based upon concepts that it defines as a "unit of thought", i.e. an idea or notion such as "shoreline emergency access" or "oil spill". Concepts may also carry other information, such as their relationships to other concepts and information about their provenance and version history. SKOS provides the means for grouping those concepts together as either collections or schemes. A SKOS collection is a grouping of concepts which share something in common and can be conveniently grouped under a common label, for example "SeaDataNet agreed parameter groups" or "ISO19115 topic categories". Similarly, SKOS concept schemes are also groupings of concepts but the relationships between the concepts are a part of the concept scheme. For example, if the eBay auction categories were published as a concept scheme, "Antiques" and "Sporting Goods" can be identified as SKOS topConcepts, the broadest definitions in the pyramids of concepts. The narrower concept definitions such as "Antique Clocks" and "Sailing" can also be delivered in the concept scheme, including their position in the hierarchy of concepts, as illustrated below. Therefore, concept schemes are a useful model for the publication of thesauri, for example the "ICAN coastal erosion thesaurus."



An illustrative example of top concepts in SKOS, and the first level of their associated narrower terms.

[21] http://www.w3.org/2004/02/skos/

SKOS also defines three forms of relationship between concepts. A concept may be broader or narrower than another concept, or related to another concept. The related attribute allows the loose mapping of one concept to another, allowing the resource to become orthogonal (see page 6). The broader and narrower attributes allow the construction of a hierarchy. If a concept belongs to a hierarchical scheme and is an entry point to that hierarchy (that is, at the top of the tree) it can be declared as a SKOS topConcept. For concepts in the same scheme, the broader and narrower relations may be said to be transitive; that is a concept two levels below a given concept can be inferred to be narrower than the concept in question without explicitly stating a relationship. For example (and illustrated below), eBay has “Sporting Goods” as a top level auction category, or a topConcept. Narrower than this is “Sailing”, and still narrower is “Rope”. If these relationships were declared as transitive “Rope” could be inferred to be narrower than “Sporting Goods”, which is not explicit in the non-transitive SKOS narrower relationship.



An illustration of transitive relations in SKOS using terms from the eBay classification of auctions.

The differences between SKOS concept collections and concept schemes are very limited in the W3C’s specification. The NETMAR project has chosen to use schemes as a discovery tool for concepts, and collections to store and publish concepts and for referencing their identifiers.

The NETMAR semantic framework has additionally extended the SKOS model to allow synonyms to be identified using the Web Ontology Language’s[22] sameAs attribute. This clearly allows the labelling of the relationship between two concepts which are identical, which is not a feature of the basic SKOS model.

Deploying ICAN semantics in the NETMAR semantic framework

Incorporating a Knowledge Organization System

The simplest way for an ICAN community member to develop a new controlled vocabulary or thesaurus (or propose new content for an existing vocabulary or thesaurus) for incorporation within the framework is to create two worksheets in a spreadsheet: one for concept names and definitions; the other for relationships between concepts.

[22] http://www.w3.org/TR/owl2-overview/

The first worksheet, illustrated below, should contain columns for

1. Concept key
    - An identifier for the concept, unique within the vocabulary. It does not need to carry any meaning.
2. Concept name and title
3. Concept alternative name (e.g. abbreviation)
4. Concept definition.

Concept Key	**Concept name and title**	**Concept alternative name**	**Concept definition**
74PQ	Plymouth Quest	PQ	{"title": "RV","callsign": "MEEU8", "platformClass": "research vessel", "commissioned": "2004-03-24","previous_name": "Sigurbjorg"}

Each concept must only occupy one row of the worksheet. If the definition needs to carry some structured information (such as information regarding the identity of a ship's hull or the bounding box of a geographic area), this should be encoded using an alternative to XML, such as the JavaScript Object Notation (JSON) standard, i.e. enclosed in curly brackets and formed of "key":"value" pairs separated by commas. For example:

> {"title": "RV", "callsign": "MEEU8", "platformClass": "research vessel", "commissioned": "2004-03-24","previous_name": "Sigurbjorg"}

The second worksheet should contain three columns describing the relationship between concepts:

1. Subject
    - The subject of the sentence describing the relationship.
2. Relationship
    - Narrower, broader, related or sameAs mapping.
3. Object
    - The object of the sentence describing the relationship.

Subject	**Relationship**	**Object**
74PQ ("Plymouth Quest")	Is narrower than	http://vocab.nerc.ac.uk/collection/L06/current/31/ ("research vessel")
74PQ ("Plymouth Quest")	Is narrower than	http://vocab.nerc.ac.uk/collection/L19/current/SDNKG04 ("platform")

Once complete, the spreadsheet should be submitted to enquiries@bodc.ac.uk along with supporting information about the domain scope of the concepts, the content governance for the knowledge organization system and the name and contact details for those authorised to make changes to the resource. The supporting information for the ICAN Coastal Erosion thesaurus, for example, is:

- Domain scope: “Thesaurus containing coastal erosion dataset (including GIS layer) terms compiled by ICAN and mapped to a global thesaurus. Includes both markup and discovery terms from the mapped components.”
- Content governance: “International Coastal Atlas Network”

The knowledge organization system will be deployed in the NETMAR semantic framework and further updates can be made by authorised persons through a web interface accessed from the British Oceanographic Data Centre website[23].

Accessing the Knowledge Organization System

Once deployed within the NETMAR semantic framework, a knowledge organization system can be accessed in much the same way as a web site, using Uniform Resource Locators[24] (URLs) to navigate the NVS. The base URL for the NVS is:

http://vocab.nerc.ac.uk

Catalogues of the SKOS concept collections and schemes hosted on the NVS can be accessed at:

http://vocab.nerc.ac.uk/collection/

http://vocab.nerc.ac.uk/scheme/

Once the identifier for an individual collections or schemes is known, it can then be accessed from:

http://vocab.nerc.ac.uk/collection/*collection_id*/current/

e.g. http://vocab.nerc.ac.uk/collection/C17/current/ is the URL for the International Council for the Exploration of the Seas platform codes collection from which the example worksheets above were taken

http://vocab.nerc.ac.uk/scheme/*scheme_id*/current/

e.g. http://vocab.nerc.ac.uk/scheme/ICANCOERO/current/ is the URL for the ICAN Coastal Erosion thesaurus

Finally, an individual concept can be accessed through this form of URL:

[23] https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_editor/
[24] http://en.wikipedia.org/wiki/Url

http://vocab.nerc.ac.uk/collection/*collection_id*/current/*concept_id*/

e.g. http://vocab.nerc.ac.uk/collection/C17/current/74PQ/ gives access to the concept definition for "Plymouth Quest" which was described in the example worksheets above

The collection URLs also provide a mechanism for accessing any concepts which have been removed from the collection (known as deprecation), or only those concepts which are currently accepted members of the collection or all the concepts which have ever been part of the collection (the default if neither deprecated, accepted or all is specified as a suffix to the collection URL):

http://vocab.nerc.ac.uk/collection/*collection_id*/current/deprecated/

http://vocab.nerc.ac.uk/collection/*collection_id*/current/accepted/

http://vocab.nerc.ac.uk/collection/*collection_id*/current/all/

The ../current/../ portion of the URLs given in this section is a shortcut to the most recent version of the collection or scheme. This can be replaced with an integer value in order to retrieve a given version of a collection or scheme.

In addition to this URL based access, application developers can make use of Simple Object Access Protocol (SOAP)[25] based access described in the associated Web Services Description Language (WSDL) document[26].

Bridging to existing Knowledge Organization Systems

Labelling data and metadata using a knowledge organization system is a first step to making those data interoperable with other datasets. However, if the knowledge organization system has defined relationships to other systems the likelihood of the metadata and data being discovered and reused alongside other data increases. Linked data is an initiative of the World Wide Web Consortium to create a web of data described knowledge organization systems. The diagram on the next page shows how this web of data is highly interconnected.

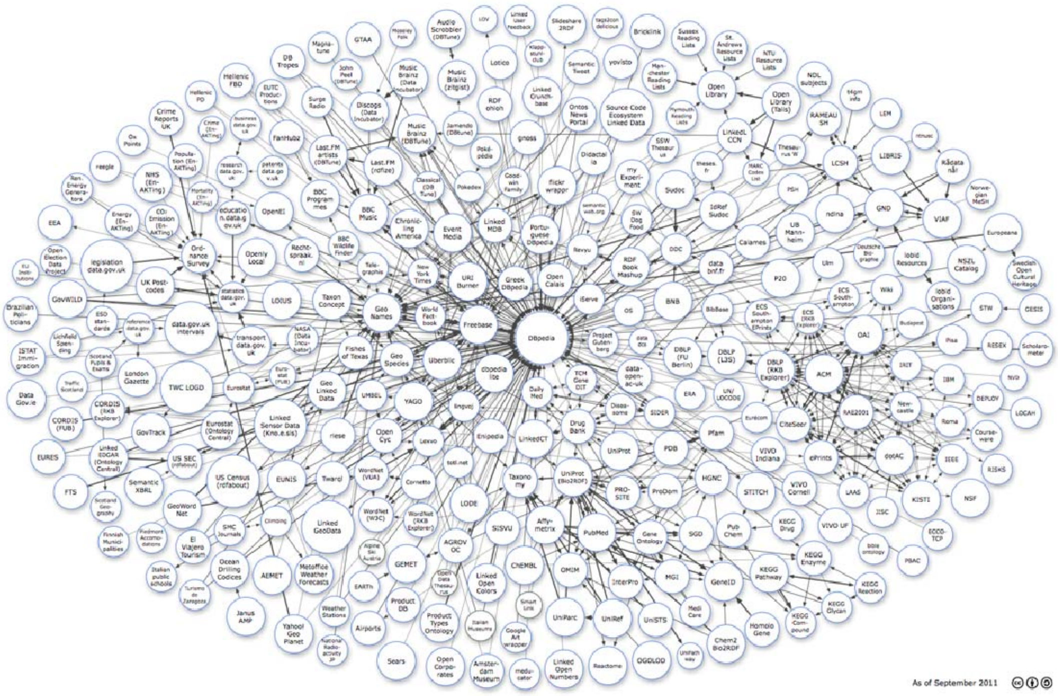
A range of environmental science and geospatial knowledge organization systems exist that may be of interest for bridging a new knowledge organization system too. These include those stored in the NVS and the Marine Metadata Interoperability Ontology Registry and Repository[8]; the European Environment Agency General Multilingual Environmental Thesaurus[12]; and GeoNames[27]. Relationships between a concept in the NVS and any external concept can be specified in the same way as the internal mappings (see page 7) but with the NVS URL replaced by the URL of the external concept as the object of the relationship. For example:

http://vocab.nerc.ac.uk/collection/P21/current/MS10360/ (sulphides)
"broader"

[25] http://en.wikipedia.org/wiki/SOAP
[26] http://vocab.nerc.ac.uk/vocab2.wsdl
[27] http://www.geonames.org/

http://www.eionet.europa.eu/gemet/concept/4350 (inorganic substances)

http://vocab.nerc.ac.uk/collection/C19/current/3_1_2_1/ (Adriatic Sea)
“sameAs”
http://sws.geonames.org/3183462/



The Linking Open Data project cloud[28].

Incorporating knowledge organization systems in ICAN metadata

This is described in detail in the accompanying cookbook: “Connecting Your Atlas.” However, in overview, the web address (URL) of a term defined in a knowledge organization system should be incorporated within a metadata document, in the appropriate field. This may be as either a string, in an XML element such as gco:CharacterString, or as a reference from an anchor field, using the xlink:href=”http://...” syntax.

Acknowledgements

This cookbook was written for the International Coastal Atlas community under the auspices of the NETMAR (Open Service Network for Marine Environmental Data) project. NETMAR is partially funded by the European Commission under Theme ICT-2009.6.4 ICT for environmental services and climate change adaptation of the Information & Communication Technologies FP7 Programme.

[28] http://richard.cyganiak.de/2007/10/lod/imagemap.html

This document has been reviewed by, and incorporates comments from, Jennifer Andrew and Roy Lowry of the British Oceanographic Data Centre; Torill Hamre of the Nansen Environmental and Remote Sensing Center; Yassine Lassoued of the Coastal and Marine Research Centre, University College Cork; François Parthiot of CEDRE; Peter Walker of Plymouth Marine Laboratory; and John Helly of the San Diego Supercomputer Center. Thanks go to the reviewers for their help in making the document clear and readable. Further feedback on this document is welcomed, and may be provided by contacting the author whose details are below.

Document Information

Author	Adam Leadbetter, British Oceanographic Data Centre	
Contact	alead@bodc.ac.uk	
Version	2.1	
Date	2012 July 26	
Revisions	2.1	Responses to NETMAR internal review. Fixing spelling mistakes; adding new definitions; clarifying some existing definitions; additional concept scheme diagram.
	2.0	2012 July 17: Comments from ICAN community with respect to discovery of existing resources included
	1.0	2011 December 19