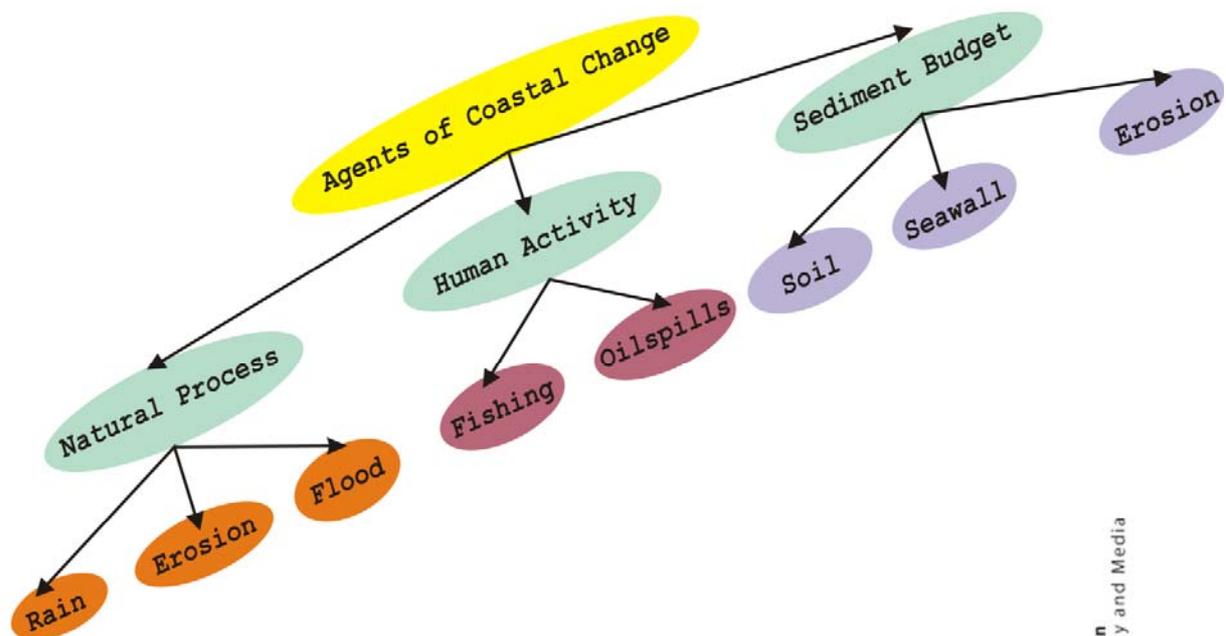


International Coastal Atlas Network Cookbook: Understanding Semantics



European Commission
Information Society and Media

Table of Contents

Introduction	3
Why use a “knowledge organization system”?	3
What are vocabularies, thesauri and ontologies?.....	4
How to define the content of a knowledge organization system?	6
Can I reuse existing resources?	6
What is the scope of the knowledge organization system?	6
Identifying the content	6
How narrow or broad should a term definition be?	6
Linking term definitions together	6
Ensuring the quality of the content of the Knowledge Organization System	7
Making the content available.....	8
Simple Knowledge Organization System	8
Deploying ICAN semantics in the NETMAR semantic framework	9
Bridging to existing Knowledge Organization Systems.....	12
Acknowledgements	13
Document Information	13

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-Lee¹

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”) accordingly. Either the mark up used throughout the system must use a common set of phrases, or there must be a means of translating between the phrases used at different points of the system, using common “semantics”. 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 scientists and data managers.

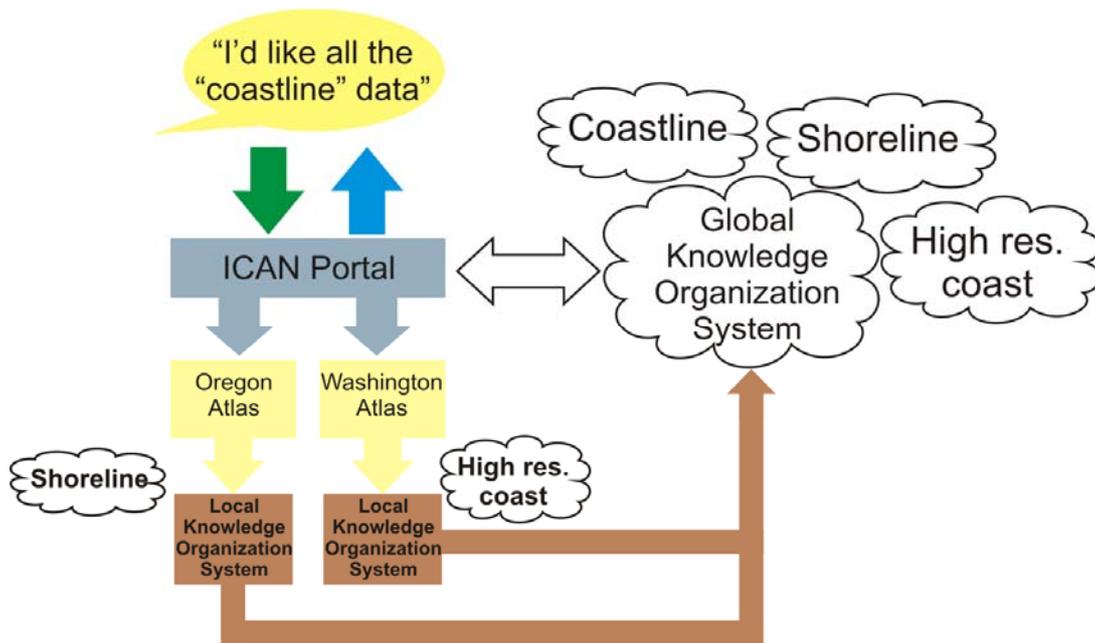
Why use a “knowledge organization system”?

One scenario for using knowledge organization systems in the International Coastal Atlas Network² (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.

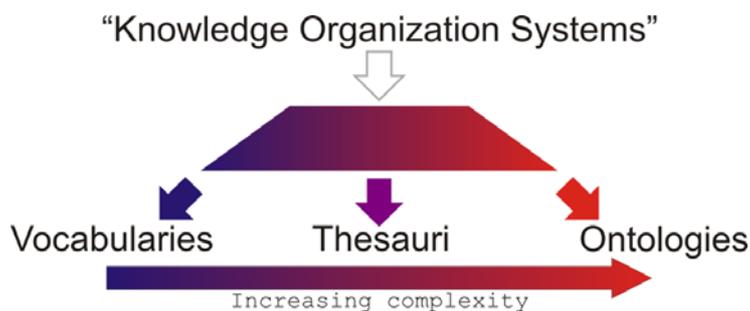


A diagram illustrating one use for knowledge organization systems in the ICAN community.

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.

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.

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. This provides the simplest form of a formal ontology. A well known example of this form is the Yahoo! web directory⁵ or the categorisation of auctions on the eBay homepage⁶. 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 Directory⁷ 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 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.

More complex ontologies can be created through declaring a term to belong to a particular class, the addition of property information to the term and the restriction of values data associated with the term may take. 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”.

⁵ <http://dir.yahoo.com/>

⁶ <http://www.ebay.com/>

⁷ <http://gcmd.nasa.gov/>

How to define the content of a knowledge organization system?

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 12). Details of how to access an existing knowledge organization system relevant to the International Coastal Atlas Network are provided on page 11 of this document.

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^{8,9}. 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.

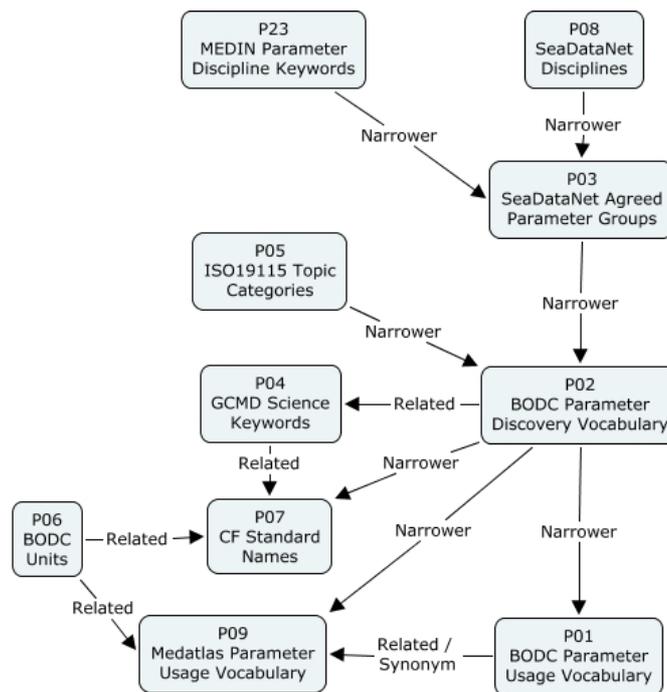
Linking term definitions together

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

⁸ 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.

⁹ 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.

knowledge. This enhances the ability of a user to find data labelled with a given 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.



An example from the NERC Vocabulary Server¹⁰ 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 Forecast¹¹ netCDF metadata conventions group; the SeaDataNet and MarineXML Vocabulary Content Governance Group (SeaVoX)¹²; and the NETMAR ontology governance body¹³. The

¹⁰ <http://vocab.nerc.ac.uk/>

¹¹ <http://cf-pcmdi.llnl.gov/>

¹² https://www.bodc.ac.uk/data/codes_and_formats/seavox/

¹³ <http://netmar.nerc.no/>

role of the content governance group is analogous to the International Organization for Standardization (ISO) definition of a “control body”¹⁴.

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 Server¹⁰ (NVS). The role of the technical governance group is analogous to the ISO definition of a “register manager”¹⁴.

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 System¹⁵ (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.

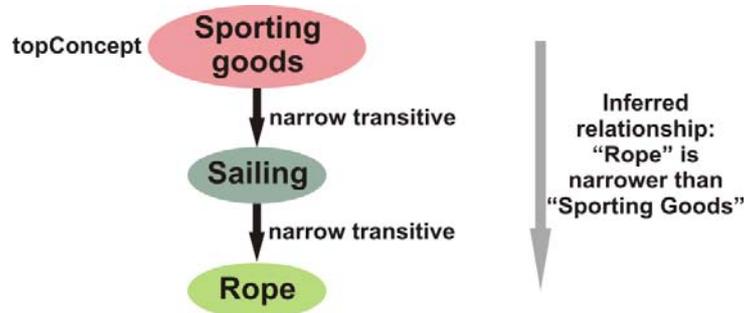
SKOS is based upon concepts that it defines as a “unit of thought”, i.e. an idea or notion such as “shoreline emergency access” or “oil spill”. Concepts may also carry other information, such as their relationships to other concepts and information about their provenance and version history. SKOS provides 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 also a part of the concept scheme, so it is a useful model for the publication of thesauri, for example the “ICAN coastal erosion thesaurus.”

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 5). 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

¹⁴ <http://www.dgiwg.org/Terminology/faq-other.php>

¹⁵ <http://www.w3.org/2004/02/skos/>

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’s¹⁶ `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.

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.

¹⁶ <http://www.w3.org/TR/owl2-overview/>

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 according 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 website¹⁷.

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 Locators¹⁸ (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:

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/

¹⁷ https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_editor/

¹⁸ <http://en.wikipedia.org/wiki/Url>

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 Repository²²; the European Environment Agency General Multilingual Environmental Thesaurus²³; and GeoNames²⁴. Relationships between a concept in the NVS and any external concept can be specified in the same way as the internal mappings (see page 6) 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"
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/
```

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.

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; and François Parthiot of CEDRE. 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	1.0
Date	2011 December 16
Revisions	

²¹ <http://richard.cyganiak.de/2007/10/lod/imagemap.html>

²² <http://mmisw.org/orr/>

²³ <http://www.eionet.europa.eu/gemet>

²⁴ <http://www.geonames.org/>