EUROPEAN COMMISSION

Community research

Climate Information Portal

CLIPC Milestone (N°: 10)

Knowledge base design

File name: { *NNN*.docx or .pdf}

Dissemination level: PU (public)

Author(s): *Wim Som de Cerff (KNMI), Peter Thijsse (MARIS)*

Reviewer(s): *NNN*.....
NNN.....

Release date for review: ***XX/XX/201X***

Final date of issue: ***XX/XX/201X***

Revision table			
Version	**Date**	**Name**	**Comments**
1	Dec 2014	Concept	

Abstract

Description of the knowledge base concept in the CLIPC project, description of the ideas and how it will be applied.

Project co-funded by the European Commission's Seventh Framework Programme (FP7; 2007-2013) under the grant agreement n°607418

Content

Introduction

The knowledge base in CLIPC is a set of services that supply explanatory information to the users when working with CLIPC services. The above schematic overview shows inputs and components of the CLIPC knowledge base.

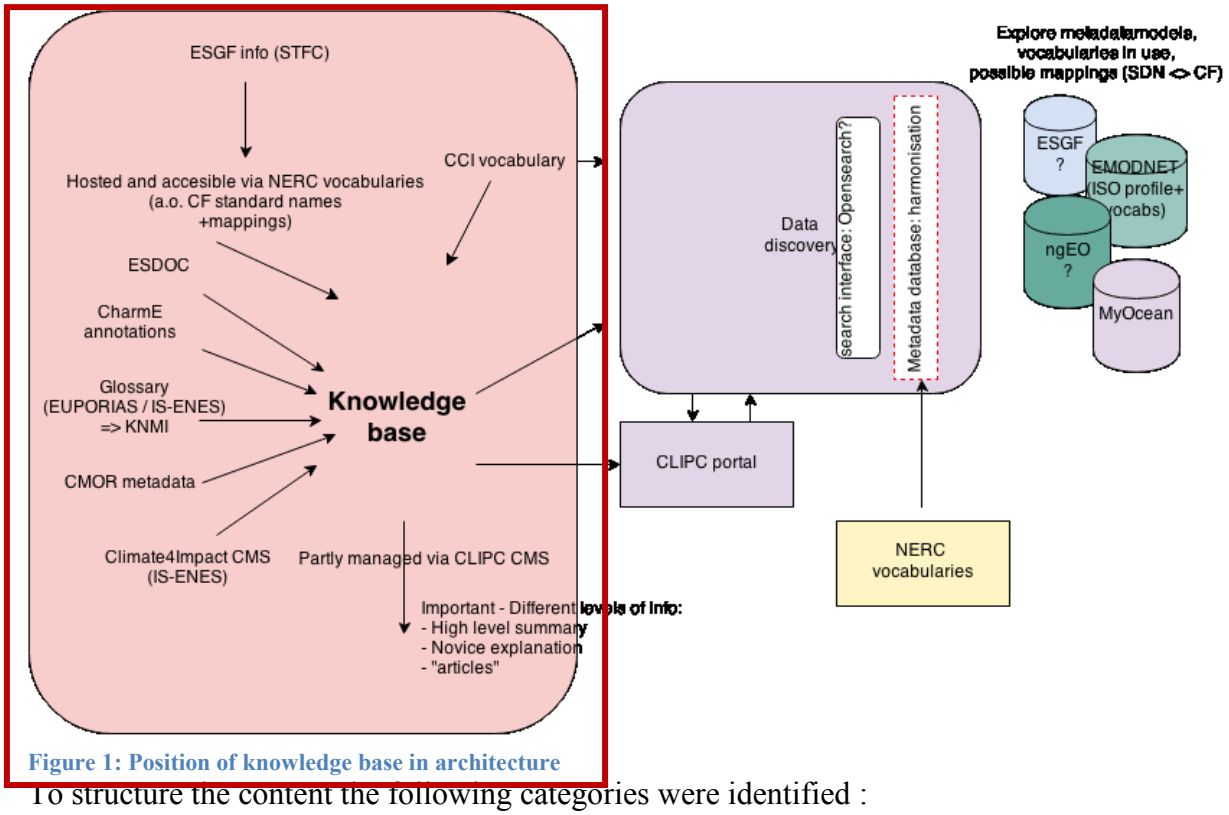

Figure 1: Position of knowledge base in architecture

To structure the content the following categories were identified :

1. Catalogue (of scientific information of datasets)
2. Commentary information
3. Technical documentation / terms
4. Glossary of terminology
5. Literature

Each category will be detailed in the next chapters.

1. Catalogue

The CLIPC catalogue of datasets will consist of:

- Scientific citations
- Author, origin
- Documentation available
- Background information
- Links to data
- Metadata model based on ISO 19115/19139

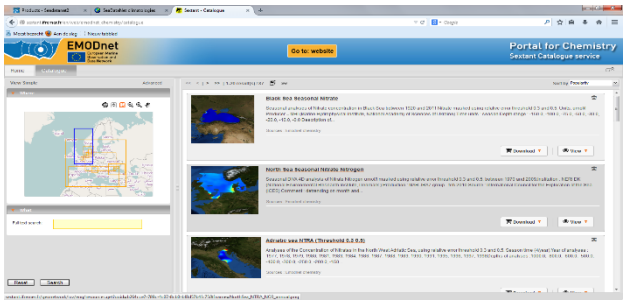

Figure 2: sample screenshot of catalogue (EMODNet Chemistry in this case)

The CLIPC catalogue will probably be developed using GeoNetwork and will be filled with validated datasets (WP6), datasets calculated/processed (e.g. in the storyline but also for other climate impact indices) within CLIPC, and metadata of climate datasets harvested from MyOcean, EMODNet and other sources.

2. Commentary information

The CLIPC portal will provide commentary information about the datasets and information that the CLIPC services offer. This commentary information is supplied by several components :

- Frequently Asked Questions (FAQ) section: This will be implemented as a set of webpages managed and updated via the CLIPC CMS. In this section issues will be explained which users often run into when using data or applications on the CLIPC website.
- Annotation to URL's to data (or better via "Handle" = group of URL's) using CharmE methodology.
  - Users provide comments on management/how to remove bad comments
  - Implementation on the portal will follow the rules of http://charme.org.uk/
  - During retrieval of dataset the CLIPC server will request the CharmE system if annotation is available.
- Version information: Version information of datasets is mainly information about a dataset provided in the CLIPC catalogue.
- Restrictions to the design of the portal:
  - Important to guide the right users to the right sections

- Add a license/disclaimer for use of portal applications and datasets: first time in a popup and message will be provided (plus always available via menu option « disclaimer ») => “When you use this website you commit to the user license and be aware that ….

3. Technical documentation – use of vocabularies

CLIPC will provide as much as possible technical documentation, explanation of terms used, and links to existing technical documentation. This part is covered by several services:
Providing definitions and documentation of the calculation and processing services implemented in the portal to generate the Tier 1, Tier 2 and Tier3 data products.
Providing definitions of the search terms in the data discovery service
Make use of the definitions and hierarchy in the SeaDataNet / NERC vocabularies
Use references to CCI documentation, which is loosely structured. Planned to import/map to BODC vocabularies via SKOS.

For the data discovery service and well as technical documentation the integration and extension of the NERC vocabulary services is a key development within the CLIPC project. A short overview will be provided in this chapter. More information on the background and use is provided in Annex 3.
CLIPC climate datasets are very diverse in origin. Although they are often already harmonised within their domain, the syntax and semantics is different. Using standardised vocabularies is a very important step in harmonised discovery and access to datasets. The NERC Vocabulary Service can assist in mapping the discovery terms to 1 single system, to optimise search and discovery.

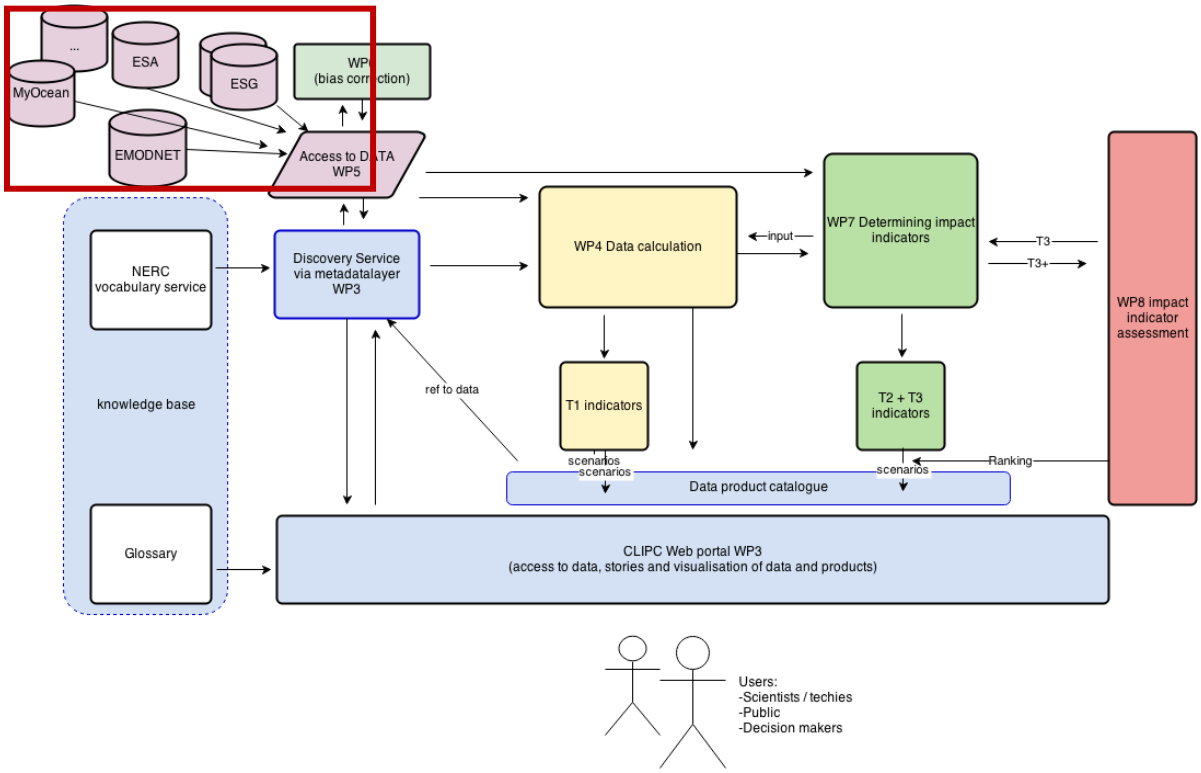

Figure 3: Position of vocabulary service to support search in various data infrastructures

Why use vocabularies? Quoting from the NERC website: *Using standardised sets of terms (otherwise known as "controlled vocabularies") in metadata and to label data solves the problem of ambiguities associated with data markup and also enables records to be interpreted by computers. This opens up data sets to a whole world of possibilities for computer aided manipulation, distribution and long term reuse.*
An example of how computers may benefit from the use of controlled vocabularies is in the summing of values taken from different data sets. For instance, one data set may have a column labelled "Temperature of the water column" and another might have "water temperature" or even "temperature". To the human eye, the similarity is obvious but a computer would not be able to interpret these as the same thing unless all the possible options were hard coded into its software. If data are marked up with the same terms, this problem is resolved.
In the real world, it is not always possible or agreeable for data providers to use the same terms. In such cases, controlled vocabularies can be used as a medium to which data centres can map their equivalent terms.
The NERC vocabulary service can be requested via:
http://www.bodc.ac.uk/products/web_services/vocab/

Summary of the NVS:

- NVS has 100+ vocabularies, plus e.g. hierarchy for discovery of parameters = discovery in steps from Disciplines to Parameter category, to individual observed parameters.

- Example hierarchy in vocabularies:
    - P07 CF standard names is part of the hierarchy scheme as are the P01 terms.
    - P07 => P02 => P03 => P08!! (every step is broader) see illustration in diagram below.

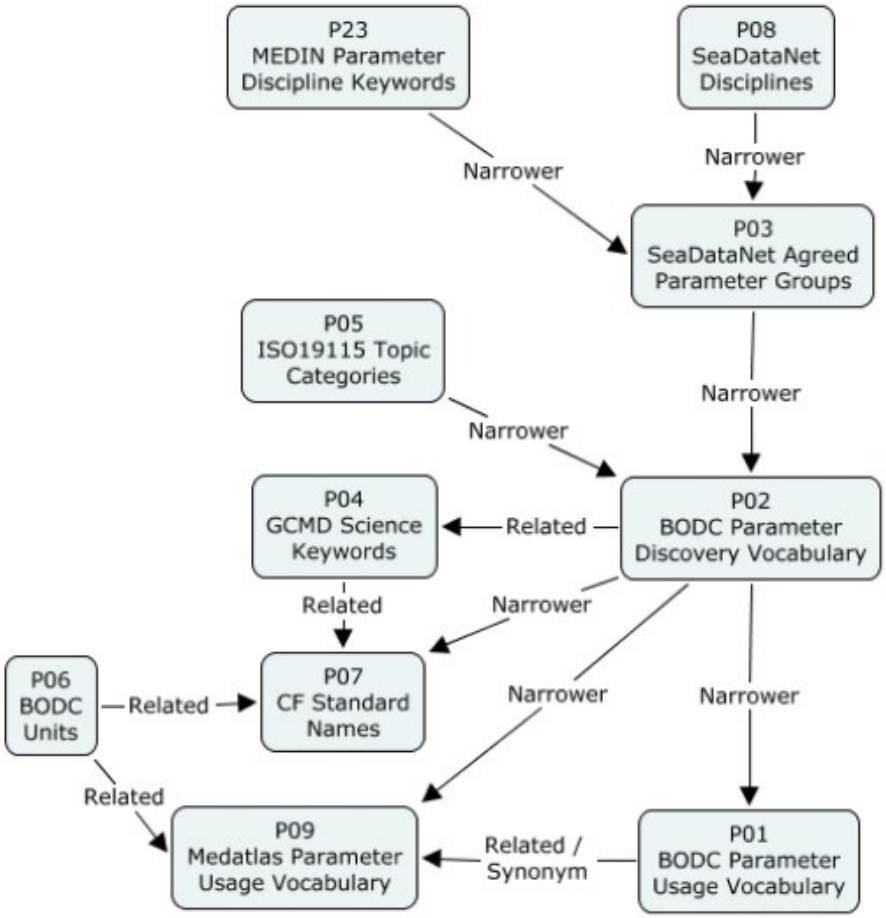

Figure 4: NERC vocabulary relations and hierarchies

The NERC Vocabulary Service will assist in mapping the discovery terms to 1 single system, to optimise search and discovery.

- CLIPC will make use of the NVS in discovery of CLIPC. Metadata terms, and especially the parameters terms, in the central CLIPC catalogue will be mapped using the vocabulary lists and mapping maintenance tools. Especially the CF <> P02 parameter mapping is very important.

  As example: P07 (CF parameter names as in use in ESGF and MyOcean) <=> P02 (in situ discovery parameters as in use in SeaDataNet/Emodnet) mapping is currently partly implemented. This will be extended in CLIPC. The CLIPC discovery service makes use of this mapping to create a specific search for a CF term, while the user on the CLIPC portal uses the P02 term.

P07 CF names is very flat, but could get some internal hierarchy implemented if requested. Example: min temp and max temp, fall under "temp", while all 3 are separate terms in P07.

- Once mapped the discovery of data based on a parameter name is more efficient users can drill down from Discipline level, to parameter group, to parameter discovery term to detailed parameter for all mapped resources. And next to this the interface can make use of the hierarchy that is available in NVS, see the visualization of the relations: http://seadatanet.maris2.nl/v_bodc_vocab_v2/vocab_relations.asp?lib=P08

Apart from the discovery support the definitions of terms will be part of the CLIPC technical documentation.

More information on the background, use and technical implementation of the NERC / SeaDataNet vocabulary services is provided in Annex 3.

4. Glossary of terminology

Next to the technical documentation also « softer » documentation about terminology will be available via several glossaries.

- The glossary created by EUPORIAS is validated and used in the IS-ENES website. It will be integrated in the CLIPC web portal and extended. EUPORIAS makes use of a Google doc as source for the Glossary and can be shared to other websites (Within the Euporias project a Drupal module was developed for this). Terms in the Glossary will get a “href-link” in the HTML webpage.
  CLIPC can develop an extra Glossary for the Climate Impact indicator terminology. The different glossaries can be used at the same time.
  Webpages as now available in EUPORIAS can be shared and included in the CLIPC webportal.

Integration of the Glossary in the CLIPC portal will be done via the RDF-a technique (underlining terms in the website, plus marking the terms in the HTML code in a specific way.). Very beneficial for Google ranking.

Example page how to apply RDFa in HTML for links to vocabs: http://www.bodc.ac.uk/data/published_data_library/catalogue/10.5285/41479c42-4dfb-4da9-be97-4c532ce13922/. There are plenty of Chrome based plugins (e.g. RDF Detective) which can extract the RDF from the page.

- CMS climate4impact ‘use case ‘ glossary: The climate4impact website contains a lot of useful documentation pages regarding use cases for climate model data : http://climate4impact.eu/impactportal/documentation/guidanceandusecases.jsp. This information will be reused in CLIPC, by using a scraping technology (dynamic) having only the Climate4Impact website as main source.