CLIPC DELIVERABLE (D -N°: AT report v1)

WP3/4/5 Architecture Team report version 1.1

File name: { docx or .pdf}

Dissemination level: PU (public)

This document describes the CLIPC architecture in concept. It is based on a concrete story line, needing real data, the required data flow around it, processing, documentation, visualization and delivery to the end user. Goal is to implement this story line with the architecture sketched in this document before end 2014 to test the concepts and stimulate development.Earlier versions of this document have been discussed with the CLIPC WP Leaders, as well as with WP7 and 8 members regarding the storyline and integration of climate impact indicator data.

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

Author(s):
Adam Leadbetter – BODC
Wim Som de Cerff – KNMI
Maarten Plieger – KNMI
Stephen Pascoe – STFC
Hasse Goosen – Alterra
Channah Betgen - Alterra
Peter Thijsse – MARIS
Ernst de Vreede - KNMI

Reviewer(s): ***Annemarie Groot, Rob Swart, Kristin Boetcher***

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

Release date for review: Jan 2015

Revision table			
Version	**Date**	**Name**	**Comments**
1.0	Jan 2015	S. Pascoe, W. Som de Cerff, P. Thijsse, M. Plieger, A. Leadbetter, Channah Betgen, H. Goosen, a.o.	First release based on earlier drafts
1.1	Feb 2015	P. Thijsse	Feedback included from R. Swart, A. Groot, K. Boetcher

Executive Summary

The objective of this document is to describe the CLIPC architecture concept. It is written by the architecture team consisting of representatives from WP3, 4 and 5.

The architecture is based upon a first real world story line, which is defined together with WP7 and WP8. The story line involves producing an urban heat vulnerability map for large European cities. The story line prototype will be delivered before the user group meeting on February 4, 2015, based on data readily available and is merely a showcase usable for discussion with e.g. user groups. When improved data becomes available during the project, the prototype can evolve into a real product. Working with a story line helps in discussions and with applying the concepts to reality which will help detecting design flaws early on. It can also be used by WP2 in discussion with end users.
The final CLIPC system will not be limited to this one storyline, but will be covering more tier 1/2/3 climate impact indicators for which the CLIPC system contains or imports the bias corrected "raw" datasets, and the processed datasets (tier 1/2/3).

CLIPC is not starting from scratch and will reuse existing concepts, components and services as much as possible. Main challenge for the architecture team is to select reusable components and services and integrate them into one architecture for CLIPC. Therefore we started with analyzing results from related projects (e.g., ESGF, Climate4impact, ClimateAdapt, SeaDataNet/EMODNET), analyzing metadata discovery standards, data access services, etc..

This first version of the document describes the architecture as an integration of 5 concepts:

- *Knowledge base*
- *Data discovery and access*
- *User management*
- *Data processing*
- *Data visualisation.*

Content

1. Introduction

The Architecture Team, with representatives from WP 3, 4 and 5, has documented a first version of the necessary architecture for the functions in the CLIPC system. This report provides the conclusions of the internal discussions, and discussions in the portal design workshop.

The document describes the approach taken (chapter 2), some background information and reference information (chapter 3), the story line (chapter 4), the architecture (chapter 5) and the open issues and discussion points (chapter 6)

The document is written to guide the development process with respect to the integration of CLIPC services, and developed in the different work packages. Other work packages are urged to make use of the conclusions of this document: E.g. WP4 for its toolkit and visualisation developments, WP3 for the conceptual design of the portal, etc.

2. Work plan

The Architecture team follows the following approach:

1. *Start with definition of story line together with WP7/8 (ALTERRA/PIK)*
   The story line will provide the basis for discussion and development.
2. *Analysis of user needs, existing concepts, services and components.*
   From the WP2 work the user community is identified. Based on the DoW and the work done in related projects existing services and components are identified
3. *Describe the CLIPC Architecture ideas: Based on the previous points, a draft architecture was set up.*
4. *Discussion with other WP's and at workshops.*
   The work done has several times been discussed with the WP leaders to verify the choices made and discuss on the next steps. At the specific portal design workshop at KNMI in November 2014 architecture and development ideas were discussed with experts from other projects (like IS-ENES, MyOcean, and other).
5. *Document the first architecture design with definition of services and interrelations (this document), after which development can start.*
6. *Development of services*
   After final agreement of the WPL's WP3, 4 and 5 can start developing the (prototype) services.

3. Background principles and definitions

This chapter describes principles and definitions of the CLIPC architecture based on the inventory of user needs, existing infrastructures solutions and required services. Chapter 4 and next chapters describe how they will be used, integrated and applied.

3.1 CLIPC architecture principles

The CLIPC architecture principles are meant as directing guidelines for taking development decisions during the development of the CLIPC services. There are two categories of principles: business- and information-layer principles. For the IT layer, no principles are defined yet.

Business layer principles:

- The architecture principles are valid for all aspects of the CLIPC project
- Services and components are loosely coupled
- The infrastructure supports the (business) goals of CLIPC
- The infrastructure is secure and reliable

Information (data) layer principles:

- CLIPC will use available components and services (re-use principle)
- CLIPC will use a service based architecture
- CLIPC will use Open Standards
- Each component has a responsible party
- Software is Open Source
- Presentation, process and business logic are independent of each other

3.2 Target user of CLIPC services

In CLIPC D2.1 an extensive survey on the target users of CLIPC can be found. Quoting from D2.1 Chapter 4.2, priority user groups for CLIPC:

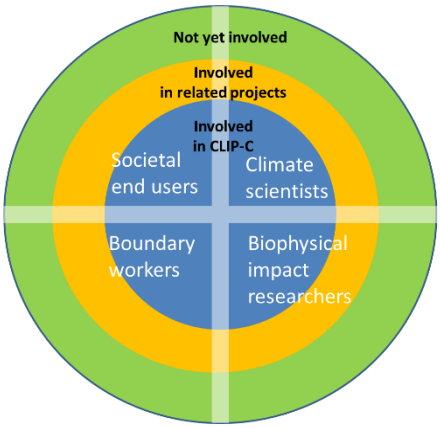


Figure 1 : Suggested classification of users groups in three circles dependent on connection to the project

First circle: Potential users already involved in projects of CLIPC partners;

Second circle: Users already involved in other similar European and national projects. The CLIPC partners participate in these related projects;

Third circle: Potential users of interest known by various partners but not necessary involved in any projects at the time.

To get a more targeted picture of the user needs we subdivide each of these three groups into four categories, according to expected requirements and capabilities to handle climate:

A. Climate Scientists

B. Biophysical impact researchers

C. Boundary workers (or intermediary organizations) and socio-economic impact researchers

D. Societal end-users

Category A involves climate scientists but also includes the data providers themselves, who for instance need observational data to evaluate their model results or to make use of empirical ground truth in other processing in order to make their results more applicable (e.g. in mapping results).
The biophysical impact researchers group (category B) may overlap with the 'climate scientists' to some extent, but this category is not involved in developing and running climate models. This group may include the downscaling community who need data for empirical-statistical downscaling or bias correction, in addition to model validation. However, most of this group will be people from the hydrology, biology, agriculture, and

engineering communities, with some experience dealing with data, and statistics. 'Boundary workers' in category C include consultants who work at the interface between the scientists and the societal end users. Boundary workers perform as intermediaries. Organizations such as the EEA but also consultant agencies and national portals can be considered boundary workers. Category C also involves socio-economic impact researchers who are assumed to apply climate data with less scientific literacy and numerical skill than the impact researchers.
The societal end users in category D represent policy makers (people involved in policy making within governmental institutions and business firms), decision-makers (people who are making the actual decisions like politicians) and practitioners (people involved in the implementation of adaptation such as NGOs, civil servants who often do not have a high climate or science literacy).

In CLIPC the focus will be on user groups A, B and C. The societal end-users will only be reached via the boundary worker, who first have to structure and represent the data products of CLIPC.

3.3 Story line

The CLIPC Architecture team has chosen to use a concrete story line as guidance in the architecture documentation and later development work. The storyline is a real world case needing real data, a data flow around it, processing, documentation and visualization. The storyline tests out the ideas and developments immediately as a first use case. The big advantage of this approach is that there are quick results in the development work in the form of a prototype, which make communication internally in the project and to external users much easier. Also, design flaws can be detected in an early stage.

3.4 Knowledge base

The CLIPC Project will develop new and incorporate (upgraded) existing specialised tools, and data services serving data and information to the users. To the various types of users of the portal it is very important to be presented with information on how to work with the tools, how to understand information/data and how to interpret the outcomes.

The development of the CLIPC knowledge base will present the necessary information to the user. In CLIPC the knowledge base will not be just one interface or service, but a combination of several services, reusing where possible existing documentation and services created in other projects.

3.5 Relevant projects

The following projects are seen as important background/reference projects for the CLIPC architecture and service development and have received special attention from the Architecture team. In this paragraph the highlights and AT comments are listed.

The projects will be distinguished in projects that have a very strong relation and will be supply services to be integrated in CLIPC, and project that are loosely related but are still kept in sight regarding developments and standards.

3.5.1 Relevant projects integrated in CLIPC

a) *ESGF (Earth System Grid Federation, esgf.org)*

Characteristics:

- Global (CMIP5), Regional (CORDEX) and comparison (MIPS) climate model data
- distributed system, searching in metadatabase of the nodes
- Some nodes replicate data from other nodes (performance, syncing then automatical. Search result contains link to dataset (NetCDF).
- Data download via Opendap. Option to view timestamp of dataset without download of complete set (comparable to NCwms)
- ESGF has an API with faceted search interface
- Uses OpenID for authentication/authorization. (Note : ESGF is its own OpenID provider)
- Uses x509 client authentication for authentication and authorization
- Documentation using ES-DOC, which can be queried using an REST API

b) *EMODNET/SeaDataNet*

Characteristics:

- EU funded infrastructure for accessing harmonized marine datasets of in-situ observations
- CDI (Common Data Index) is the ISO19139 INSPIRE compliant metadata profile to describe the datasets. In the metadatacode lists as managed in the NERC vocabulary service are used as much as possible. Dataset transport format are mainly ODV ASCII and NETCDF (SDN profile or CF).
- NERC Vocabulary service support uniform semantics in metadata, and provides faceted search facilities: http://seadatanet.maris2.nl/v_cdi_v3/browse_step.asp
- Central portal overarching distributing datacenters

- Central portal provides metadata overview, ordering and access facilities, but download is direct action from datacenter to user.
- For interoperability with other infrastructures WMS, CSW, and OAI-PMH services available.
- Current developments under EMODNET and SeaDataNet2 are a.o. dataproducts on aggregated datasets (in various EMODNET lots), as well as for MyOcean (common temperature and salinity product for climatology)

c) *Climate4impact*

Characteristics:

- Climate4impact uses the ESGF infrastructure
- Good example of integration of ESGF and setup of system.
- Goal of Climate4impact is to provide Global/Regional climate model data to impact scientists
- Provided services (downscaling, indices calculations) can be reused in CLIPC
- Provides visualization and processing services to other portals
- Uses same security mechanism as ESGF
- Integration services are important for the "techie" scientist/developer, but should be hidden for the CLIPC user groups (CLIPC aims at broader, non-climate-science user community)

3.5.2 Relevant projects, more loosely related

a) *Climate-ADAPT portal*

Characteristics:

- Closely related to CLIPC (goal of CLIPC is to provide indicator services also to the Climate- ADAPT portal)
- Link to map viewer inside the portal: Presentation less aimed at public/decision makers, less explanation. In that sense the Klimaat effect atlas (next) fits better for CLIPC website.
- Keep in mind as background and important to interact with

b) *Klimaateffect atlas (Knowledge for Climate)*

Characteristics:

- Good example of browsing Tier3 scenario's with in the end a map interface with explanation of the scenario and access to Tier2 and Tier1 data
- Important for CLIPC to wrap this into a smooth webinterface, in order to inform and communicate to non-technical users (policy maker).
- Lesson for CLIPC: Inform the user with a complete scenario of an indicator, direct and use the underlying maps (use the visualisation made anyway in CLIPC – selection of total offer), followed by a reference to data download. The scenario integrates all!

Knowledge for Climate: http://knowledgeforclimate.climateresearchnetherlands.nl

4. Architecture of CLIPC system

4.1 Introduction

This chapter describes the foreseen architecture of the CLIPC system as basis for developments.

The CLIPC architecture will be described in a set of component related viewpoints and their integration:

- Data discovery and access
- Data processing
- Data visualisation
- Knowledge base (with a.o. use of a vocabulary service)
- User management

The overall architecture is shown in Figure 2

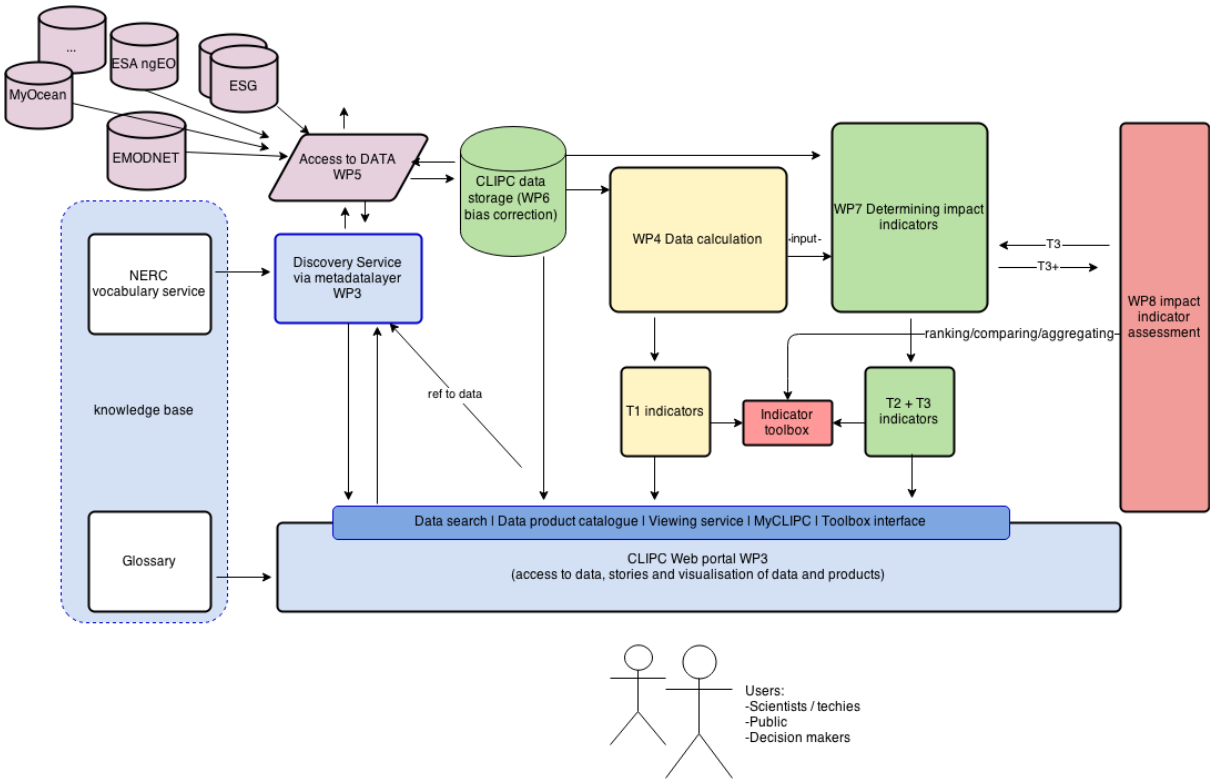


Figure 2 Overall Architecture

One of the main challenges of the CLIPC portal is to provide harmonized discovery, followed by smooth access to various sources of climate data. Resources for this data are in first instance:

- ngEO (link via partner Magellium)
- EMODNET/SeaDataNet
- ESGF
- MyOcean.

The final target are the datasets identified in the dataset inventory of D5.1.

4.2.1 Discovery service concept

CLIPC will facilitate the users in two ways to discover data and data products that are important for retrieval of climate data and information:

- Via the CLIPC data and data product catalogue
- Via a structured data search on selected infrastructures.

The content of the two services is different and they are aimed at different users. The CLIPC data product catalogue is aimed at providing validated datasets in an easy way to impact researchers and boundary workers, while the “raw” data search aims mainly at the climate

scientist (although he will also be interested too in the content of the catalogue). The CLIPC portal will have to guide the right users to the right service.

More details can be found in the next chapters.

4.2.2 CLIPC data and data product catalogue

The CLIPC data and data product catalogue is the key tool to provide information about the CLIPC datasets. It will provide the users access to the datasets that are validated and processed in the project in order to create the climate impact indicators:

- The bias corrected data from within CLIPC (WP6)
- Processed tier 1, 2, 3 datasets

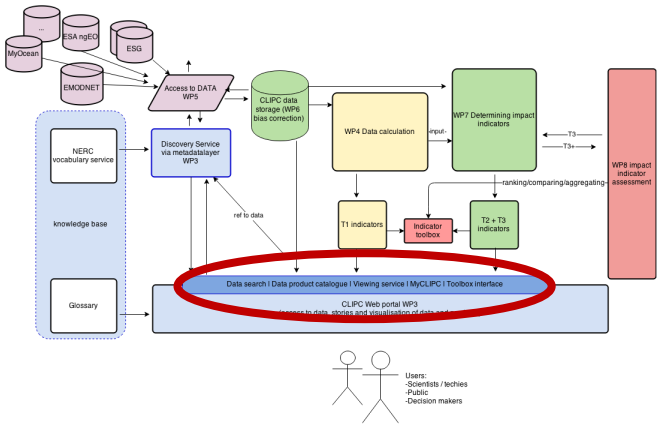


Figure 3: Position of catalogue

The catalogue has two main functions :

- Users can search the catalogue, view metadata details and get links to downloads and visualize the datasets.
- Datasets found in the catalogue can also be selected to be used in the MyClipc processing toolkit (see 4.3) where the user can start processing datasets.

The catalogue will be built up with ISO19115 / ISO19139 metadata (profile and editor to be selected) and will initially be small and mainly consist of the datasets related to the first storyline, but will be expanded later on when more climate impact indicators are created. All resulting datasets will be available in the catalogue and can be visualised by WMS/WFS services via a viewer.
The datasets can be put in a basket and be processed in the toolkit (based on Climate4Impact developments). The results can be used again by the user in the visualisation services that are available in the portal. (See chapter 4.3)

Browsing the catalogue is open for every user, but when the user want to select a dataset for processing and add it to his 'basket' the user has to login first. (More on user identification see chapter 4.6)

4.2.3 CLIPC "raw" data search

Climate data consists of satellite, in situ, re-analysis and climate model data. The CLIPC data discovery service aims to provide the users of 'raw' climate data (climate scientists mostly, but model data also for impact researchers) with an interface to search data from various connected data infrastructures.

Figure 4 shows a diagram of the discovery service to achieve harmonised search actions triggered by the user (bottom) to the diverse data resources (top). Discovery of data will be facilitated through a faceted search interface that triggers a search harmonisation layer that transforms the search actions to a request that fits the infrastructure (both technically as well as semantic).

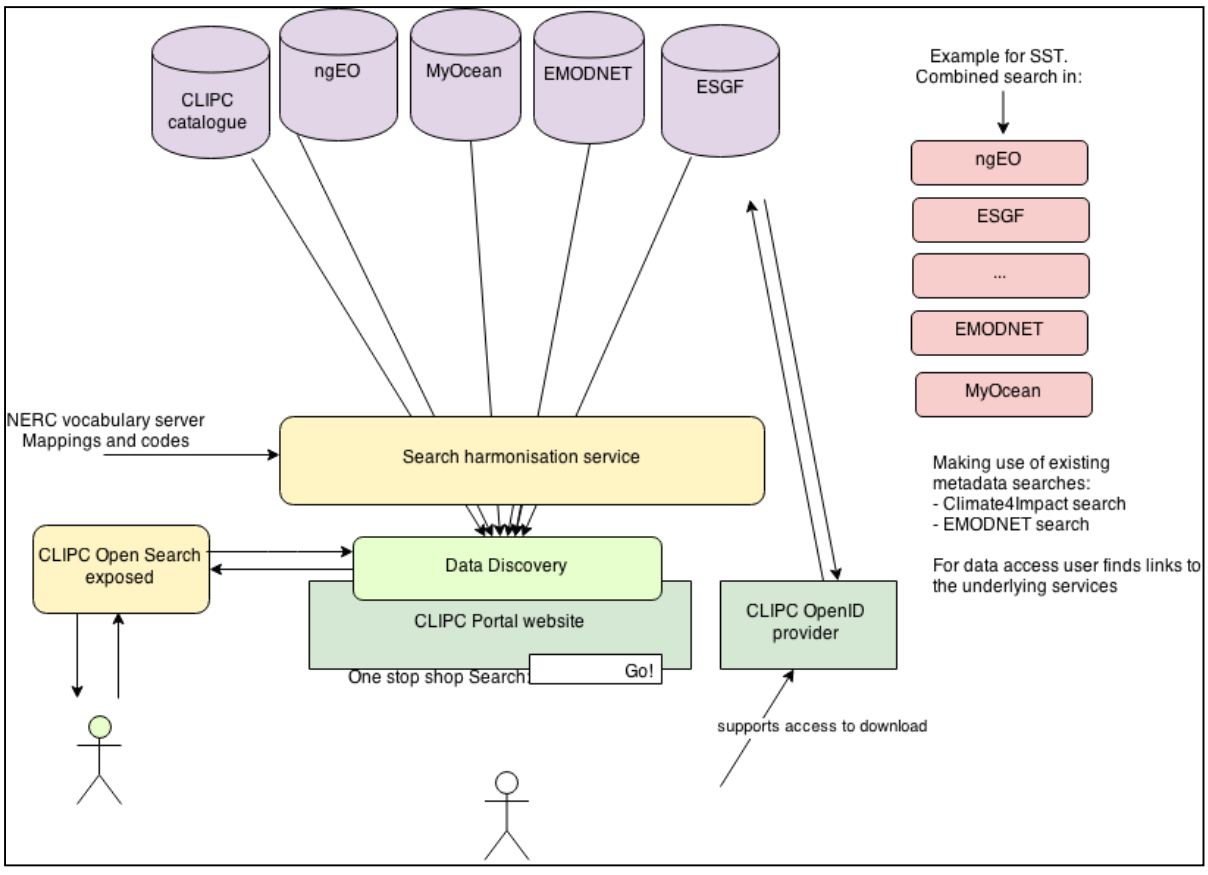


Figure 4: discovery service and security

How will the discovery service be implemented:

- The search API's (or URL's of services) of all infrastructures will be registered in a catalogue.

- The metadata model of each infrastructure is analysed and common search fields are extracted. After first research the expected common fields are: Geospatial, parameter (CF standard, or SeaDataNet standard), date/time, free search.

  The semantic model per infrastructure is different, especially for the parameters but the NERC Vocabulary Service (full specification in appendix 3) will support in the required mappings.

- The search query on the portal is translated to the required API (or webservice) call per infrastructure. Where needed the NVS provides mappings that are used on the fly.

- The response per infrastructure is not combined in one list but presented in numbers per infrastructure. This prevents that a service with millions of datasets takes too much exposure in the results, and sets from other infrastructures cannot easily be found.
  When clicking the datasets from a certain infrastructure the user is directed to the specific portal where he will find the same results, with more details, and can order and download the data via that portal.

- Additional to the dedicated CLIPC search interface also a CLIPC OpenSearch interface (linkage to overarching systems) will be developed.

- Security is an issue when moving to data download. CLIPC will reuse OpenID providers (ESGF, Google, …) to assist the CLIPC users on the CLIPC portal, and secure specific datasets. However it will focus on a Single Sign On experience for the users. When the user finds datasets in the discovery services, he/she has to login according to the policy of that infrastructure. If the ESGF infrastructure accepts the CLIPC OpenID provider, then logging in is quite easy.

- The NERC vocabulary service supports the “query harmonisation process” by supporting a mapping of the various controlled code lists into one (the one from NVS). Plus, the NVS provides for some code lists (e.g. for parameters) a hierarchy of vocabularies to support discovery from discipline level to the detailed parameter term. The metadata layer contains minimum metadata per resource plus links to the download of files or download system of the resource. More details in specific section about NVS.

The following wire frames in figure 8 show the initial ideas for the search interface on the CLIPC portal.

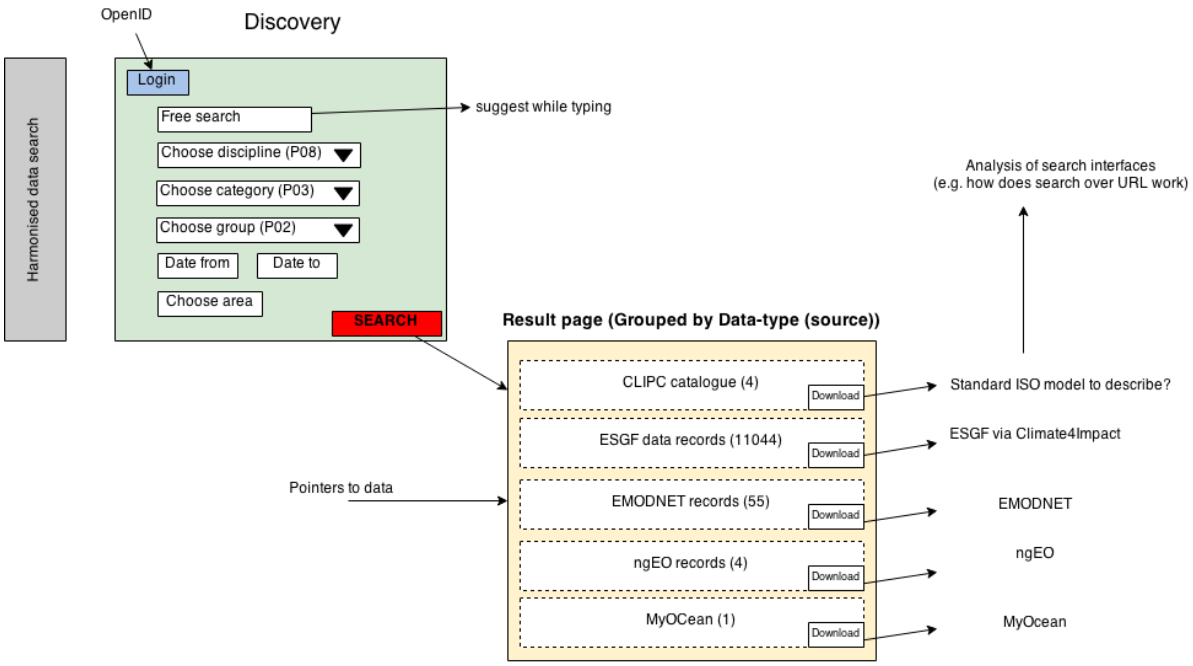


Figure 5: Sketches regarding wireframes for searching data supported by vocabularies

Example specification of the interface (fig. 4)

- Via a search interface with options for a search by parameter (group/category via NVS), data and geographical area users find a range of datasets that fulfills their requirements.
  - Free text search possible on parameters, aim to use auto-complete by using the P02/CF list of terms used in the resources
  - Daily/weekly build up index with terms in use.
  - Build up mapping for parameter codes from CMOR > CF > P01. CMOR <> CF is already available, as is CF <> P01. (ESGF allows search for CF term)
  - Other search fields:
    - Unfold of tree for P08 > P03 > P01/P02
    - Date
    - Data type: gridded / points timeseries observation
    - Remove area search (probably not usefull with many global products/datasets)
  - Not necessary to log in for the Search interface

- The dataset results are grouped per resource and supply deeplinks to the download systems of each resource. Data itself will not be buffered, nor will the metadata. However, the search query will be « harmonised» which means the query will be translated into a query matching the demands of the infrastructure (e.g. translating/mapping the parameter to the code accepted). The user is in all cases directed to the data source for downloading datasets (and retrieving metadata about quality etc) making use of the system offered. Downloading might be direct or via a registration/ordering process.

Important points related to the technical set-up of the harmonised search:

- CLIPC will NOT harvest metadata in a "buffer" but make use of existing search API's/interfaces and translate the query on the CLIPC server into queries that are suitable for a certain infrastructure.
- ESGF has a search API that can be accessed by the search harmonisation layer (e.g. search of CF parameter is possible). ESGF search is based on SOLr which uses replication rather than harvesting as its distribution mechanism. Therefore, although it is possible to query ESGF search nodes on demand, it may be (will be tested) advantageous to replicate the ESGF SOLr indexes within the CLIP-C system for efficiency. More information see Appendix 2.
- SeaDataNet/EMODNet metadata are based on the CDI metadata model : http://www.seadatanet.org/Standards-Software/Metadata-formats/CDI Metadata are exposed via WMS/WFS, CSW or OAI-PMH services available for a query to their resp. aggregated data and dataproducts. More information see Appendix 1.
- MyOcean products are available in Geonetwork CSW: http://www.myocean.eu/web/69-myocean-interactive-catalogue.php where metadata is in ISO19139 protocol and semantics make use of vocabularies. Access to data via registration.
- The CLIPC data and dataproduct catalogue (holding validated datasets, Climate impact indicators T1, T2, T3) will also be queried bij the data discovery service.

4.3 Data processing

In CLIPC there will be both pre-computed products (indicators) and a processing service. The processing tools will serve both to support science staff to maintaining pre-computed products and to provide users with the capability to explore. The exploration will be in a MyCLIPC “playground” area of the portal and is not to be confused with the pre-computed products that will have more authority.

Allowing the CLIPC users to process datasets and calculate climate impact indicators (tier 1, 2 and 3) is an important CLIPC development. CLIPC will provide an environment in which the user can select validated datasets, launch calculations and processing to the datasets, and afterward visualise the results. CLIPC will not start from scratch but make use of methods and techniques that have investigated and tested under Climate4Impact.

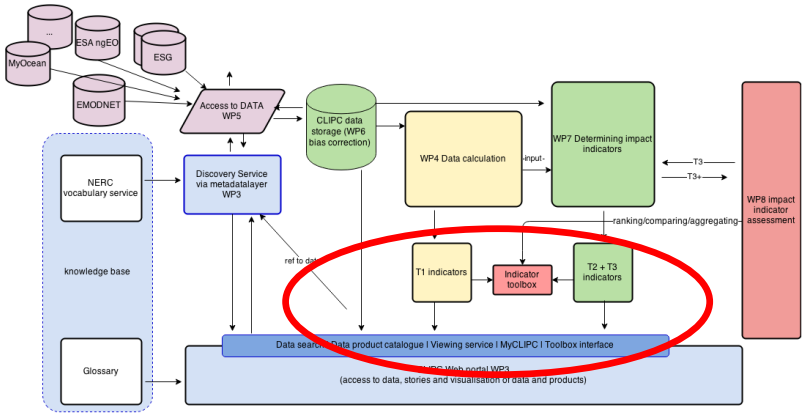


Figure 6: Data processing in CLIPC

4.3.1 Data calculation

CLIPC will provide (following KNMI developments in Climate4impact) a graphical user interface for processing. See figure 6.

The user has access to a « basket of data » and can select which files to use. The user can select files from the CLIPC catalogue to his basket, plus upload files downloaded from other infrastructures (e.g. via the CLIPC discovery services) to his own « playground » basket.

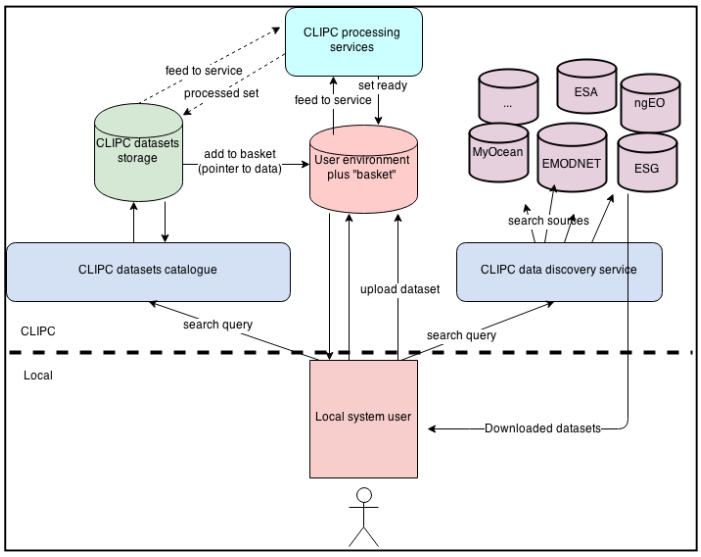


Figure 7: Use of user environment / basket

After that the users select the calculation function, plus what settings to use for the calculation. When a calculation is

started, the user can monitor progress from any location, e.g. at office or at home. Calculated products are stored again in the personal user space of CLIPC, and the user can login from any location to take a look (using CLIPC portal viewing serrvice) and download the results. Results are only publicly published when users allows this, login is always required to gain access. It is possible to re-use intermediate results in new processing to create higher level products.

More specific example: Tier1 and tier 2 calculation of climate indices can be achieved using the processing functionalities of the CLIPC portal. Climate4impact has developed some basic calculation mechanisms which will be used as basis. It is for example possible to calculate the number of Tropical days on any CMIP5 dataset containing tasmax. The processing is based on the OGC Web Processing Service (WPS) standard and uses the PyWPS framework. The service is accessible as a webservice for other portals. It is secured using the same security mechanism as used for ESGF (OpenID or x509).

The basis for the calculation of valid climate impact indicators has to be validated and bias corrected datasets. Therefore only the bias corrected (and other selected) datasets as created/selected under WP6 will be include in a catalogue of source data.

4.3.2 PyWPS

PyWPS is an open source implementation of the OGC Web Processing framework. Processes are modular within PyWPS, they are simple individual scripts. Any process or algorithm that can be described with a set of input and output parameters can be used. PyWPS does not care about what the process is doing, as long as the process gives periodically status information. This can be used to display progress information to the user. The actual process remains a black box for PyWPS. Any tool, like NCL, CDO, R or an executable can be used. In climate4impact this mechanism is used to provide simple indices calculation on datasets in the ESGF.

PyWPS can also be used for calculation framework for the Tier 2 and Tier 3 indices. It is to be discussed if Tier 3 calculations are to be made interactive or adjustable for the CLIPC researchers. This will determine if PyWPS implementation work for T3 is justified. Current viewpoint is that only Tier 1 and Tier 2 indices will be processed via the PyWPS framework and Tier 3 indices will be not be interactively calculated but « Precooked » by the CLIPC. Interactivity could be provided in the final visualisation, created by CLIPC.

4.3.3 Access to services and calculated data

Services offered by climate4impact can be accessed by other systems or portals. Climate4impact provides the same security mechanism as ESGF uses to provide access to its services. For browsers OpenID is used, for command line/system access X509 certificate public key infrastructure is used. This is the same mechanism as used in ESGF. A myproxy certificate generated by one of the proxy servers at an ESGF index node can be used to gain access to climate4impact services.

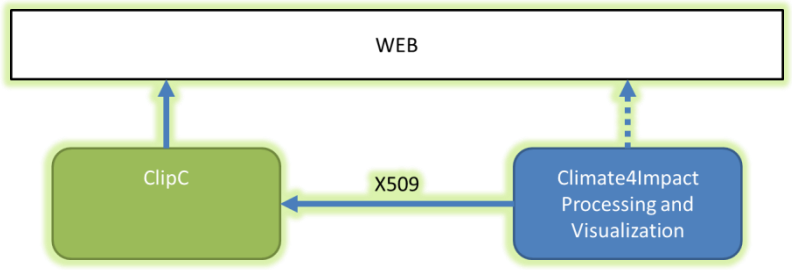


Figure 8: Possible collaboration of CLIPC using climate4impact processing capacity

The CLIPC services can be used to do initial generic computations, like Tier 1 indices. The user selects in portal environment (after login) which datasetsfrom the catalogue he wants to use for the calculation from a « basket of data ». Calculated datasets for a specific users will be published/saved in the catalogue of the user, but not made public. From his private catalogue the user can visualise his datasets using the CLIPC portal viewing service, as well as create new calculations, visually compare etc.
On the other hand official CLIPC datasets, calculated datasets by CLIPC researchers to create climate impact indicators (e.g. the sets related to the storyline) will be publicly available in the catalogue to every user.

Development approach:

- Processing services as well as data storage will be done on the climate4impact servers (running at KNMI). The ADAGUC mapservice will generate the map files of the processed datasets.
- The portal catalogue, user basket and visualisation service will run on MARIS servers as part of the CLIPC portal.
- The development will start small, only focussing on the selected datasets and processed datasets (for tier 1, 2 and 3), plus the processing services related to the storyline. All services will be developed and tested for this test case, and afterwards extended for the other storylines and with the ability for users to upload their own datasets.

4.4 Data visualisation

In WP4 and WP7 data will be processed and modelled, after which the set will be visualized using specific software making use of OGC standards for interoperability.

The CLIPC portal will have its own central viewing service (MARIS development) to allow users to go from browsing the catalogue or basket, to viewing the datasets via available mapping services.

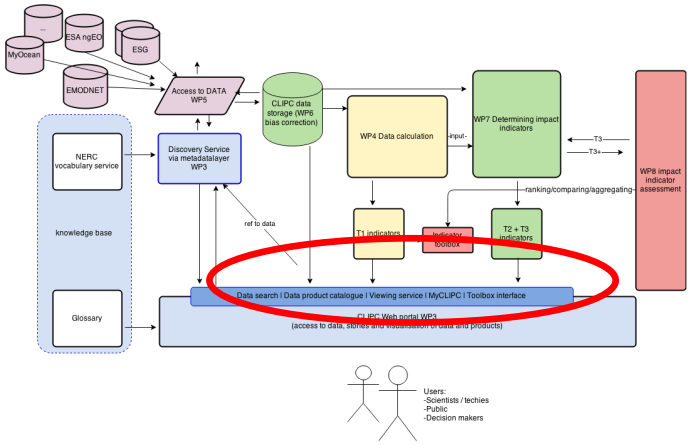


Figure 9: Location of visualisation

The viewing service will also allow to combine the (processed) model output from WP4 and WP7 (KNMI map services), with the maps from external mapservers.

4.4.1 Map generation (server side)

Goal of data visualization is to provide interactive maps to the map viewing service on the CLIPC portal. For each datasets created in the processing of datasets (calculated datasets, tier 1 – tier 3) a map (WMS/WFS) can be created when requested.

For creating these maps existing tooling can be used. In general, Tier1 products are less complicated to build/calculate than Tier3 products. Calculation and visualization of model data and Tier1 products will make use and build further upon development in the climate4impact portal (see 4.3). Maps and data offered by the system can be combined with other geographical data from other selected resources (external WMS/WFS services) into a new dataset. It is possible to export data to dedicated desktop GIS packages and tools like InDesign to enable manual fine tuning. Tier3 products are complicated to build and require manual expertise from specialists to provide meaningful results before they can be published e.g. as interactive maps.

The steps required to go from raw datasets to Tier3 is shown in Figure 10 Process of designing an interactive map for CLIPC.

Step 1	Calculation of Tier 1 indices. Tier 1 calculation can be achieved using climate4impact technology. The results can be visualized, products can be made suitable for e.g. InDesign. Visualizations, results and processing from

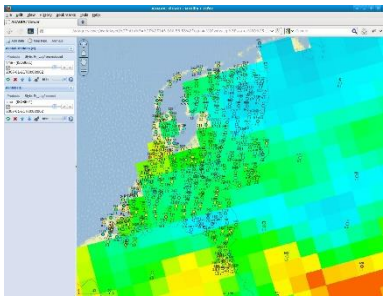
	climate4impact can be made accessible directly to the CLIPC portal. Data can be converted using Web Coverage Services to any projection and format, allowing data to be used in dedicated design tools like Adobe InDesign and ESRI ArcGIS.
Step 2	Calculation of Tier2 products (only possible for Tier2 indicators without complex models) . This can possibly also be accomplished using climate4impact technology: Tier1 products can be re-used and combined to form Tier2 products. Tier2 requires combination of several data sources with different origin and format and is more complicated than the creation of Tier1 products. Therefore only selected datasources can be used, otherwise it will lead to erroneous products.
Step 3	Tier3 products. Highly specialized product, is made by specialists and cannot be automated. Interpretation and integration of Tier2 products and statistics. Maps and graphs are made using Adobe InDesign. End products can be interactive PDF documents or customized web applications.
Step 4	Publish maps on CLIPC portal.

Figure 10 Process of designing an interactive map for CLIPC

Please note: Not all visualization work can be fully automated. It requires work of specialists and interaction with users to create these informative visualizations.

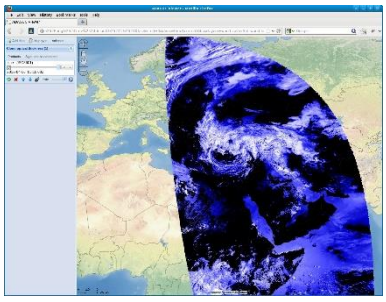
4.4.2 ADAGUC as mapserver

ADAGUC is a geographical information system to visualize NetCDF files or OpenDAP data resources via the web. The software consists of a server side C++ application and a client side JavaScript application. Within CLIPC mainly the server side will be used. The software provides several features to access data and generate mapped data over the web. It uses widely adopted OGC standards for data dissemination.

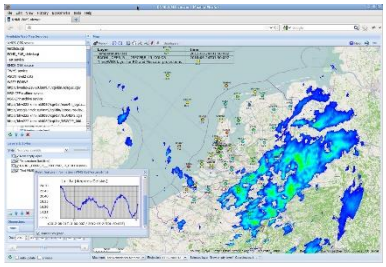


Observations

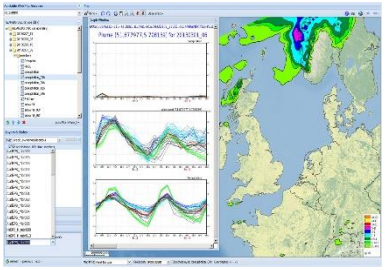
Radar



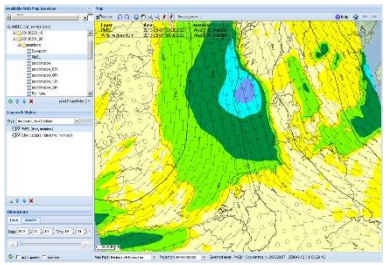
Satellite



Timeseries



Ensembles



Meteorology

The spatial data infrastructure is based on OGC compliant web services. These web services include Web Mapping Services (WMS) for online visualization, Web Feature Services (WFS) allowing downloading vector data and Web Coverage Services (WCS) for downloading raster data.

Interoperability with other WMS systems

ADAGUC consists of a client server architecture which is tightly coupled, but the system keeps its interoperability with other WMS services. The server and client can be used standalone as well. Data served with the ADAGUC server can be visualized in any WMS compatible web based tool like OpenLayers, Leaflet or GoogleMaps. For desktop applications GoogleEarth, ArcMap and QuantumGIS can be used.

Multidimensional data with time and elevation components (t, z, x, y)

Data with a time component and/or elevation component (pressure in Pa or height in meters) can be online selected and visualized. ADAGUC is designed to work with N-dimensional datasets, it can also visualize datasets consisting of several model members or ensembles. It keeps full interoperability with the WMS standard at all times. The data itself needs to comply to the Climate and Forecast (CF) conventions. Most of the datasets offered in the ESGF system do comply to these conventions.

Application of ADAGUC in http://climate4impact.eu/

Within the IS-ENES project, ADAGUC is applied in climate4impact to provide quick looks and previews of CMIP5 data and CORDEX data. The software has been adjusted to work with ESGF data and the ESGF security system.

At climate4impact users have their own personal space where results of calculations can be stored. The datasets in their personal space can be accessed and visualized using ADAGUC.

Maps served with ADAGUC are accessible in the Klimaateffect atlas (Knowledge for Climate).

ADAGUC will be used as core mapserver of the CLIPC system to create maps from calculated and processed datasets. ADAGUC will be requested from the portal viewing service via WMS, WFS and WCS protocol.

4.4.3 CLIPC viewing service

Users on the CLIPC portal will have access to a viewing service based on OpenLayers 3. This service can combine maps that the user generates in his/her « user playground » (calculated datasets, tier 1 – 3), plus other maps from selected datasources. The datasets in the CLIPC catalogue can also be visualised via the CLIPC viewing service, and combined in the same user session.

OpenLayers 3 (http://openlayers.org/) is fully OGC compliant and can therefore visualise maps generated from WMS, WFS or WCS services. MARIS will make use of the latest version of this software as basis of the viewing service, and will combine it with Javascript and php code to develop a userfriendly interface. Below an example of such an interface. The CLIPC viewing service interface will be specified in more detail in the official deliverable D3.1 Conceptual Design of the CLIPC portal.

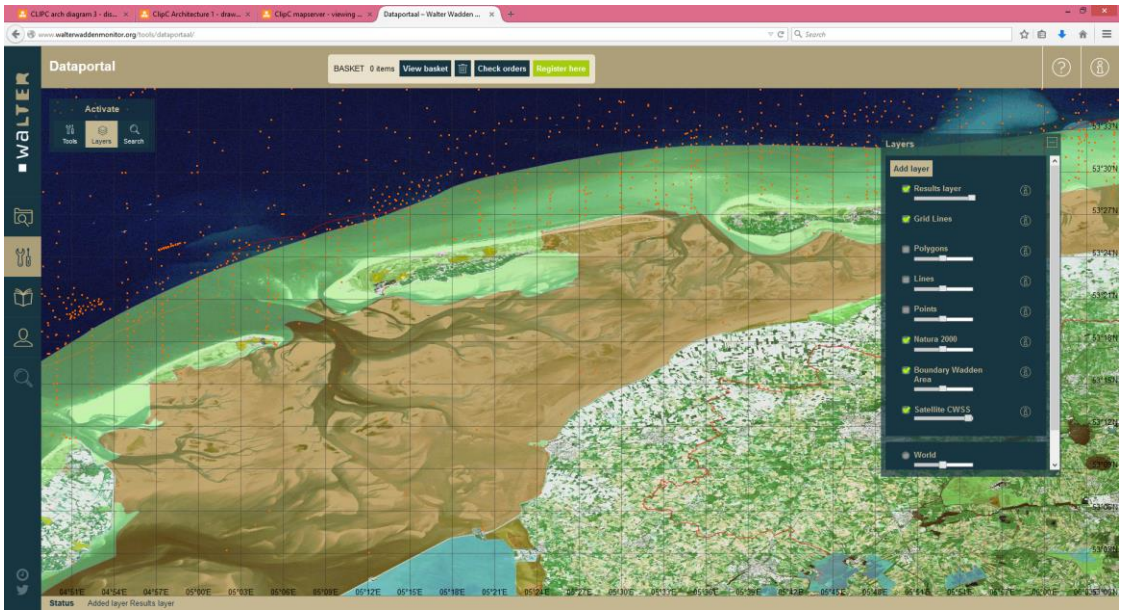


Figure 11: Sample of viewing service interface

4.5 Knowledge base

The knowledge base in CLIPC is a set of (distributed) 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.

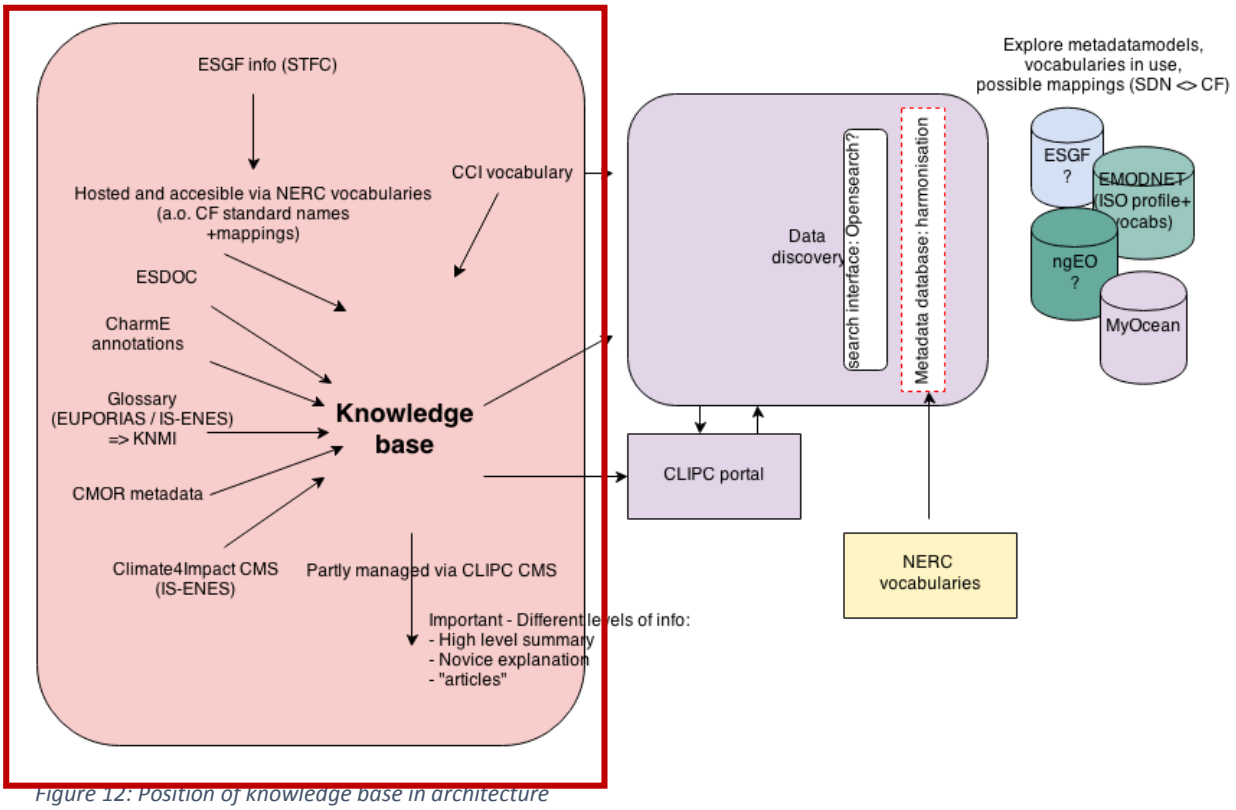


Figure 12: Position of knowledge base in architecture

To structure the content the following categories were identified :

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

Each category will be detailed in the next chapters.

4.5.1 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

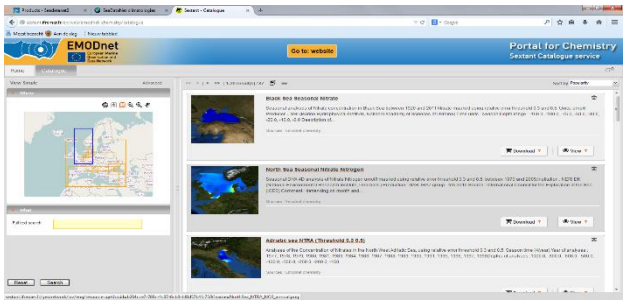


Figure 13: sample screenshot of catalogue (EMODNet Chemistry in this case)

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.
ISO profile and metadata editor to be discussed.

4.5.2 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 ….

4.5.3 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.

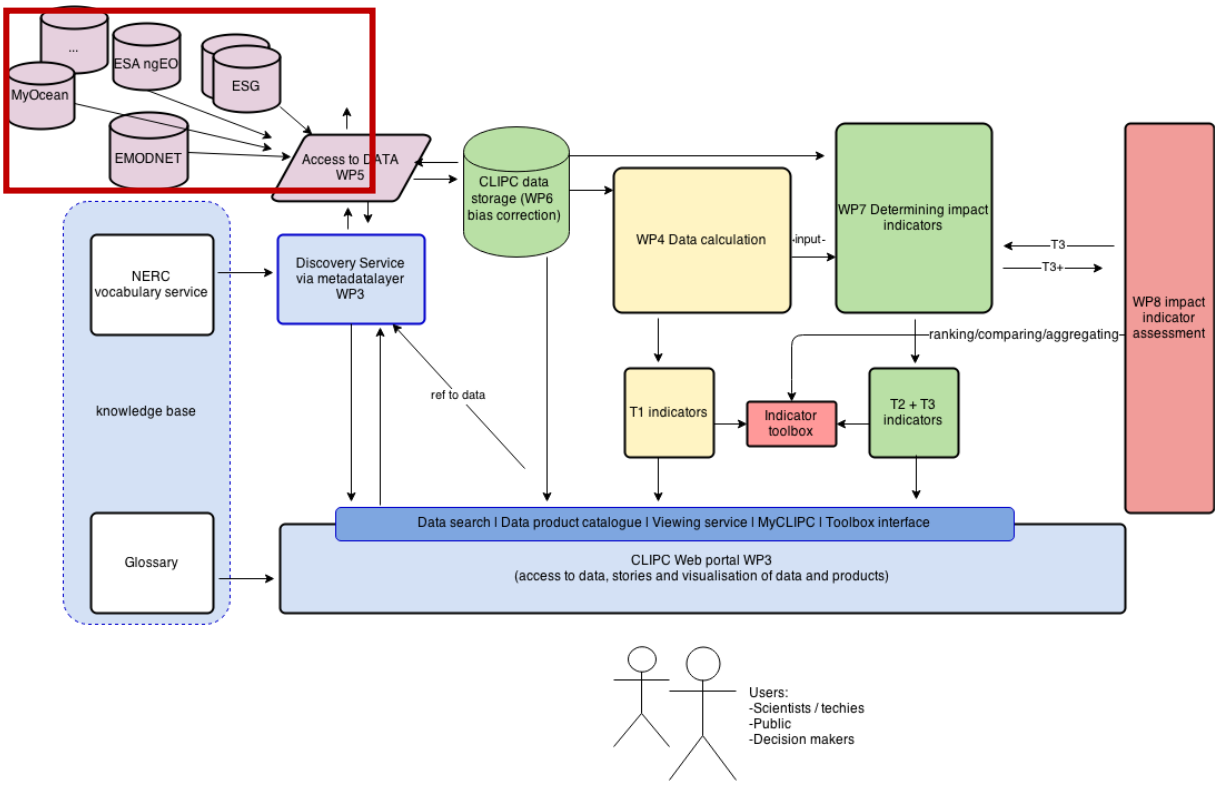


Figure 14: Position of vocabulary service to support search in various data infrastructures

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.

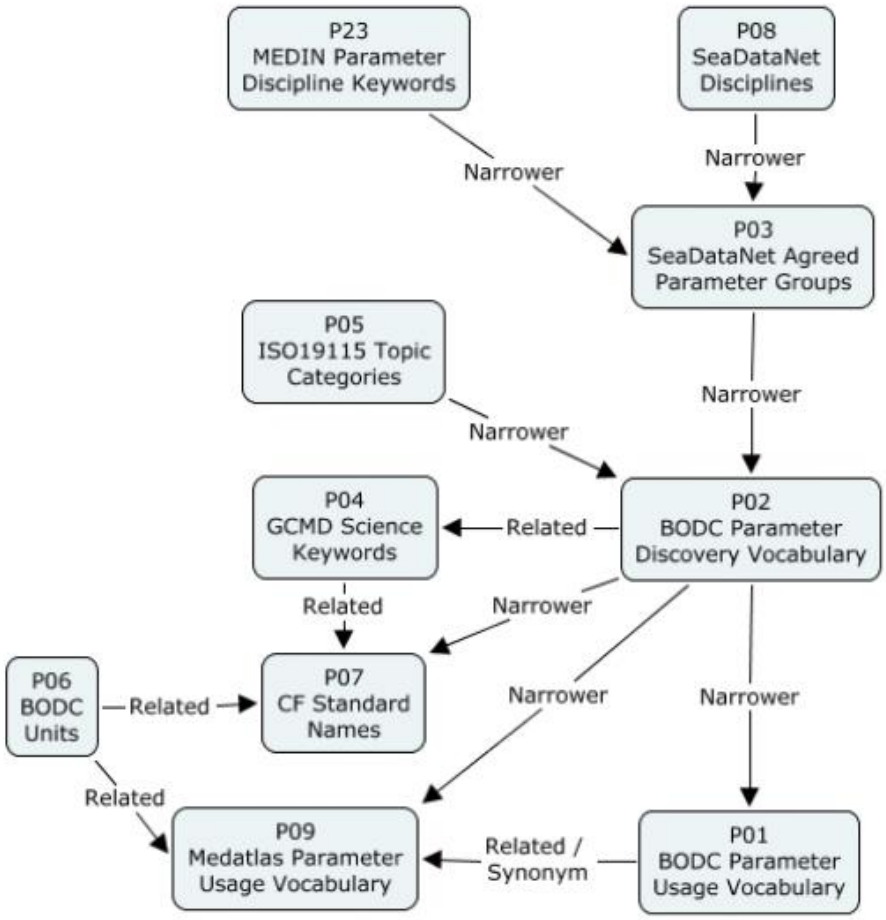


Figure 15: NERC vocabulary relations and hierarchies

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.

4.5.4 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.

4.6 User identification

Discovery of data via the CLIPC data discovery service and well as browsing the catalogue (plus visualisation) is open and no user registration or login will be needed. Only when the user is accessing the CLIPC processing services, « the user playground » the user need to login to the CLIPC services. This user identification is required and functional, because the system will have to know who the user is to provide him/her the correct content in the « basket » of datasets, also when returning to the system later.

User authorisation will be provided in two ways :

- Accepting OpenID users from selected OpenID providers such as Google, LinkedIn, ...
- Accepting OpenID users from the ESGF authorisation system, which is a second implementation.

Following the design already in use with the Climate4impact portal, the CLIP-C portal will allow users to authenticate with an OpenID obtained from an ESGF (or Google/Linkedin) identity provider node (IdP).

In this way the CLIPC system will only have a very light user management, providing as little barriers to the user as possible (only functional.).

5. CLIPC storyline: urban heat vulnerability

The storyline is used as basis for discussions about the CLIPC architecture and will be the pilot during development testing out all developed components : Processing datasets (T1,T2,T3),Discovery of datasets in the catalogue, visualising datasets, adding to user basket, adding viewing service, etc.
The flow of data is shown in Figure 16: Storyline data flow marked with red arrows.

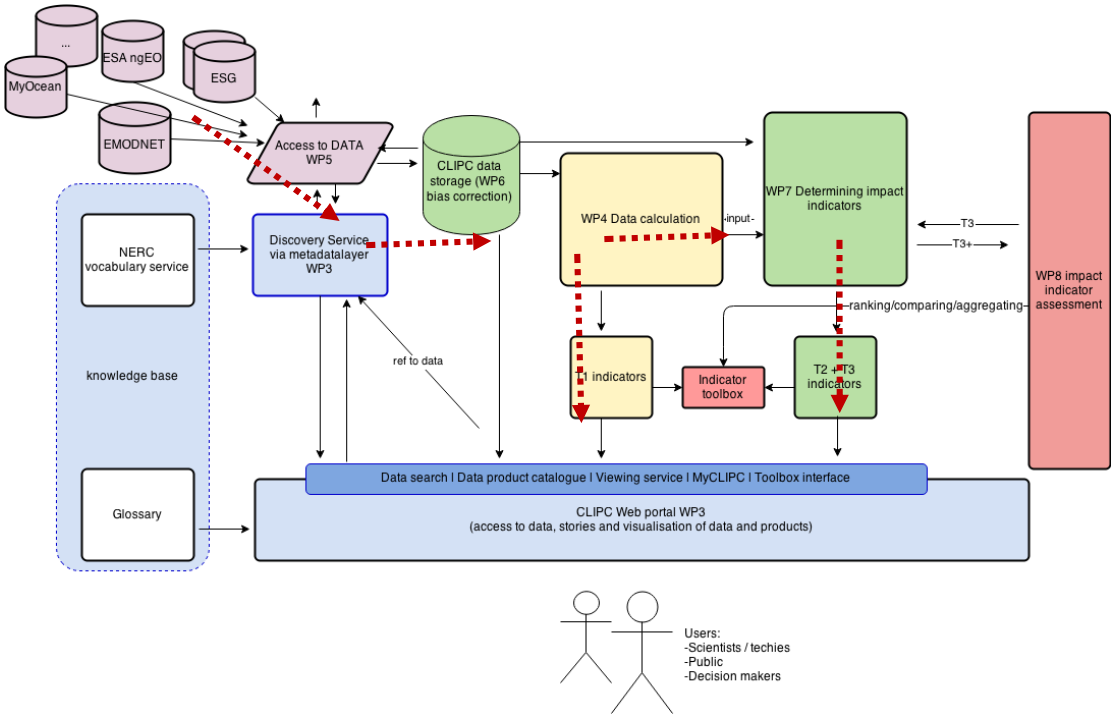


Figure 16: Storyline data flow marked with red arrows

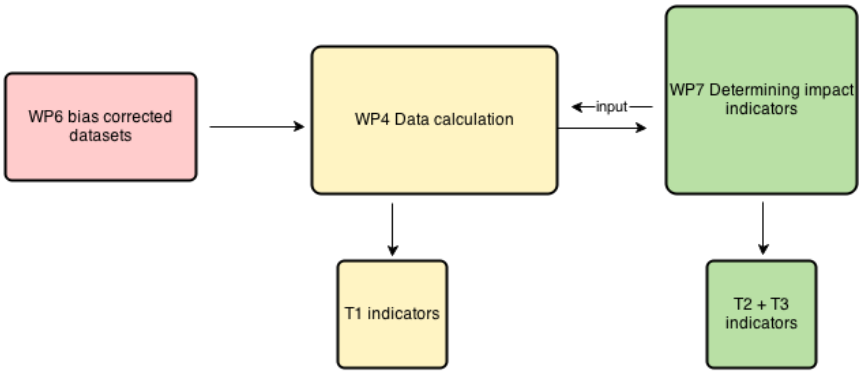


Figure 17: Storyline separated

5.1 Goal for the story line

The goal of the storyline is to provide an urban heat vulnerability interactive map for Europe (T3 indicator) interesting for policy maker at city council level (WP2: Category C: Boundary worker). Interactivity will be in comparing current and future climate scenario's for cities in Europe on different maps.

The goal of the product is raise awareness and to get the policy makers attention for this problem. The map will enable policy makers to compare between cities. The map will not provide information for local city problems because of the map resolution, more local research for the specific city is needed to answer questions on a city scale.

5.2 Datasets required (Only bias corrected data – from WP6!)

For the generation and processing of the final product the following datasets and calculated datasets are necessary.

Tier 1: temperature climate dataset: daily minimum and maximum temperatures:

- Daily minimum and maximum temperature
- Current climate (current, 30 year average)
- Future climate scenario's (which RCPs are most suitable?)
- Resolution: at least 3km^2
- Derived indices (adjustable for specific area's): the average number of days a year, where the night temperature is not lower than 20 degrees Celsius. In this calculation the threshold of 20 degrees should be adjustable (e.g., for South Europe it could be 25 degrees Celsius)

Calculation: use Climate4impact services for indices calculations using as basis only the bias corrected datasets from WP6 supplied via internal OpenDAP services.

Tier 2: Land use data:

Combine T1 with land use characteristics like thermal characteristics of cities

- Land use map of Europe (should contain at least: height, % green, %buildings/roads)
- Surface albedo map for Europe (CORINE dataset?)
- Land use scenario's for future land use in Europe
- Building surface fraction
- Impervious surface fraction
- Green surface fraction
- Element height
- Sky view factor

Calculation: Use the WUR urban heat island effect formula.

Tier 3: Sensitivity data:

Combine with how humans react to heat stress (comfort/health issues) - PIK Potsdamm can help here?

- Possibility to make use of green areas
- Population density
- Total population
- Share of the population older than 65
- Demographic dependency
- Share of people with lower socio-economic status (poorer housing quality and lack of air-conditioning)

Calculation: ?

5.3 Visualizations

The T3 visualizations as a product for boundary workers to present to policy makers cannot be fully automated. Automatic visualization of T1, T2 and T3 data for science use is possible and will be possible via the CLIPC viewing service. Generation of these maps for further appliance in creating the T3 product for policy makers will also be possible but will need to be checked, and e.g. combined in a graphical (InDesign and other) interface before presented to the public. Klimaat effect atlas is a good example visualization of the T3 data for policy makers. Alterra and PIK will take the lead in this.

List of frequently used abbreviations and acronyms

ArcMap - Main component from ESRI ArcGIS software stack

ADAGUC - Atmospheric Data Access for the Geospatial User Community (OpenSource OGC WMS/WCS implementation) – http://adaguc.knmi.nl/

CMIP5 - Coupled Model Intercomparison Project - Phase 5

CORDEX - Coordinated Regional Downscaling ExperimentCCII

CSW - Catalogue Service for the Web (ISO standard for access to catalogue data)

EMODNET - European Marine Observation Data NETwork (can be seen as data access infrastructure, mostly built upon SeaDataNet standards)

ESGF - Earth System Grid Federation - http://esgf.org/

GeoServer - Open Source map server software

InDesign - Adobe design software http://www.adobe.com/nl/products/indesign.html

Leaflets - Open source viewing service code set

MyO – MyOcean, FP7 project for marine/ocean modelling.

NcWMS - Software able to create WMS (graphical map) from NetCDF files

ngEO – Next Generation Earth Observation ESA

OGC - Open Geospatial Consortium

PyWPS - Open source python implementation of the OGC Web Processing Service standard.

QuantumGIS - Also known as QGis, Opensource light GIS application, especially for viewing graphical maps

OAI-PMH - Catalogue access standard, forthcoming from library domain.

OpenID - Open source and widely used user management service

OpenLayers - Open source viewing service code set, see Openlayers.org

SeaDataNet - Pan European marine data infrastructure focussing on metadata and data standards for marine data.

UMN Mapserver – Map server software developed by Minnesota University

WCS - OGC Web Coverage Service

WMS - OGC Web Mapping Service

X509 - Public key infrastructure (PKI), security mechanism used in ESGF

Appendix 1: Integration of SeaDataNet/EMODNet data

SeaDataNet as core part of EMODNet exposes several services for accessing the aggregated metadata (aggregated per datacenter per discipline)

- SeaDataNet OAI-PMH: http://seadatanet.essi-lab.eu/gi-cat/services/oaipmh
  SeaDataNet CSW: http://seadatanet.essi-lab.eu/gi-cat/services/cswiso
- SeaDataNet WMS:
  http://geoservice.maris2.nl/wms/seadatanet/cdi_v2/seadatanet?service=WMS&request=GetCapabilities
- Sample of WFS:
  http://geoservice.maris2.nl/wfs/seadatanet/cdi_v2/emodnet/hydrography?service=WFS&version=1.0.0&request=getfeature&outputformat=gml3&typename=points&maxfeatures=10&bbox=17.2793103448275872,40.4448275862069,17.7206896551724127,40.944827586206895
- For data access the user is directed to the central catalogue (exactly showing the datasets as found via the services) where the order can be placed and data can be downloaded.
- Data access is distributed via registration but for certain communities quick access to aggregated datasets is available after agreement.
- OpenSearch is available as well.

Appendix 2: Integration with ESGF

The ESGF infrastructure is composed of a globally distributed network of nodes which cooperate to provide a unified set of services. Data is held at one or more data nodes but can be discovered from any node in the federation. Similarly a single user management system provides unified registration, authentication and authorisation.

Each ESGF node is configured to act as 1 or more types:

1. **Data** nodes offer download and subset services. They enable download over HTTP and potentially GridFTP. They offer subsetting of NetCDF data over the OPeNDAP protocol.
2. **Index** nodes offer search services. They enable free text and facetted search using the SOLr search backend. Index nodes can query other nodes on demand to return results for the whole federation or can replicate records from other nodes.
3. **Identity Provider** nodes (IdP) offer registration, authentication and authorisation services. The OpenID standard is used to enable single sign-on across the federation. There is a group permissions system allowing users to register to access to particular datasets. These access restrictions are enforced by security components on the data nodes. Scriptable authentication is via short-lived SSL client certificates with plans to add support for OAuth2.

From the perspective of CLIP-C we can think of ESGF providing 4 types of service in a form which abstracts away the location of data: identity and permissions management, search, download and subset. Figure 9 illustrates how CLIP-C portal can integrate with these services to provide access to ESGF data. The Climate4impact portal has successfully demonstrated that visualisation can be built on top of the ESGF subset service and how a 3rd-party portal can use ESGF IdP services for login.

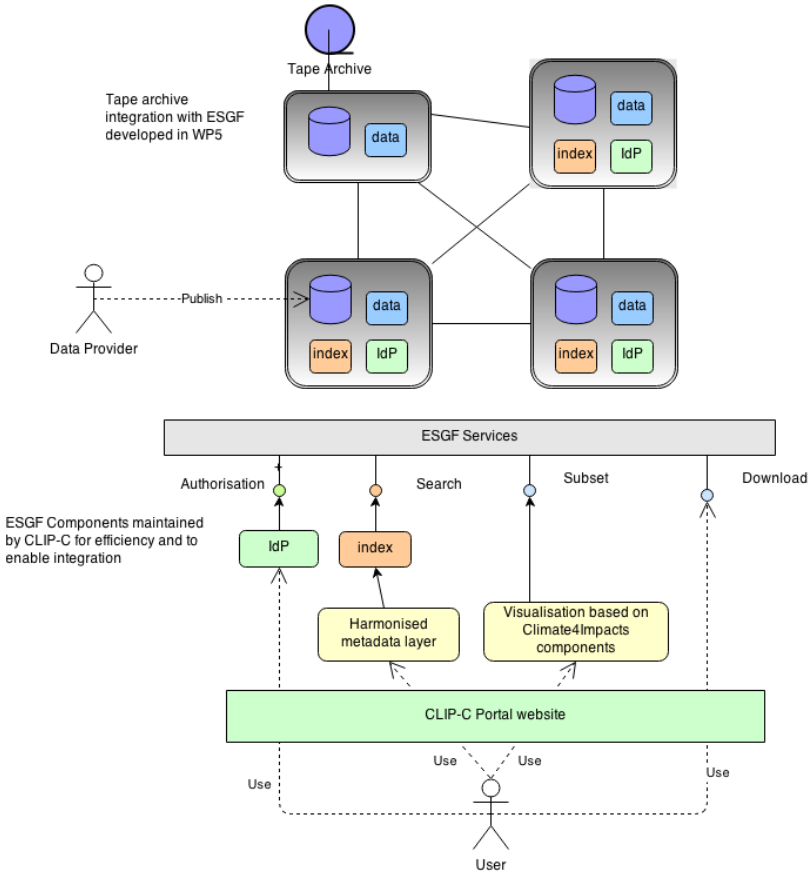


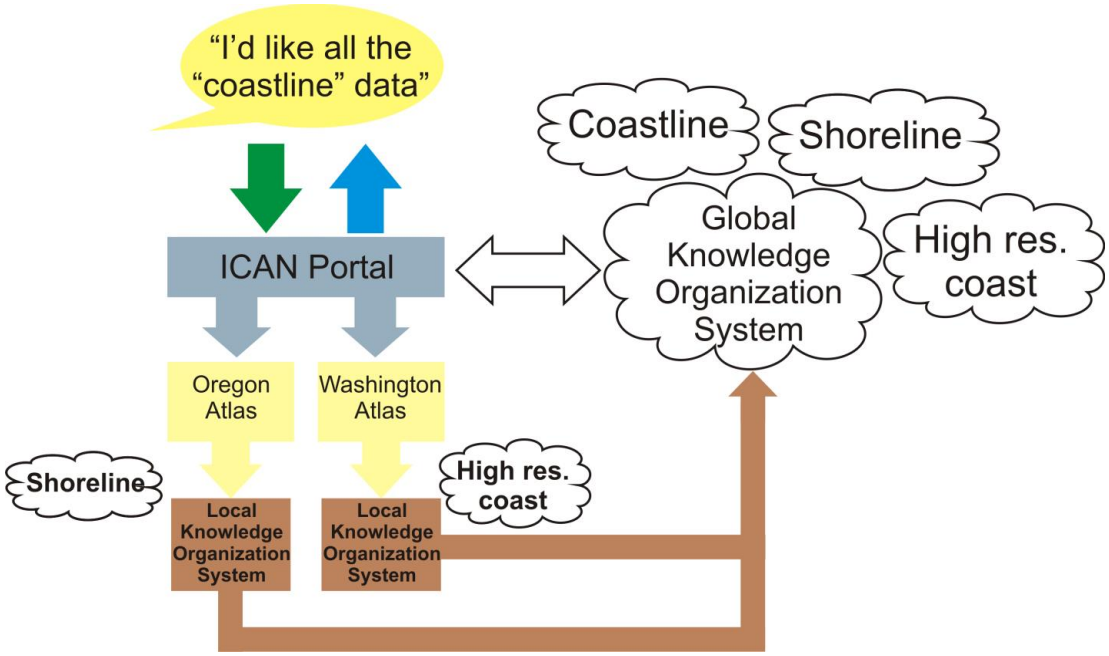
Figure 18: Harmonised data access via integration with ESGF components.

Appendix 3 : Vocabulary Services and Knowledge Organization Systems for CLIP-C (A. Leadbetter - BODC)

Contents

Why use a "knowledge organization system"?

One scenario for using knowledge organization systems is to search through the local data catalogues for a given data keyword from a central portal. For example, as illustrated below, a user arrives at the portal and requests "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"[1].



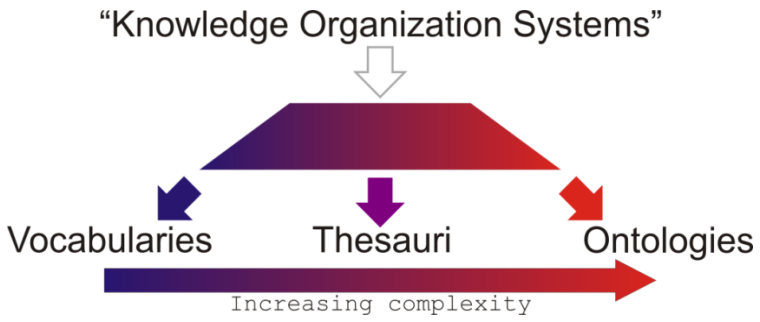
A diagram illustrating one use for knowledge organization systems.

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.

[1] 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.

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)[2].

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[3] or the categorisation of auctions on the eBay homepage[4]. 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"

[2] 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.

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

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

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[5] 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.

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 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.

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

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[6,7]. 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.

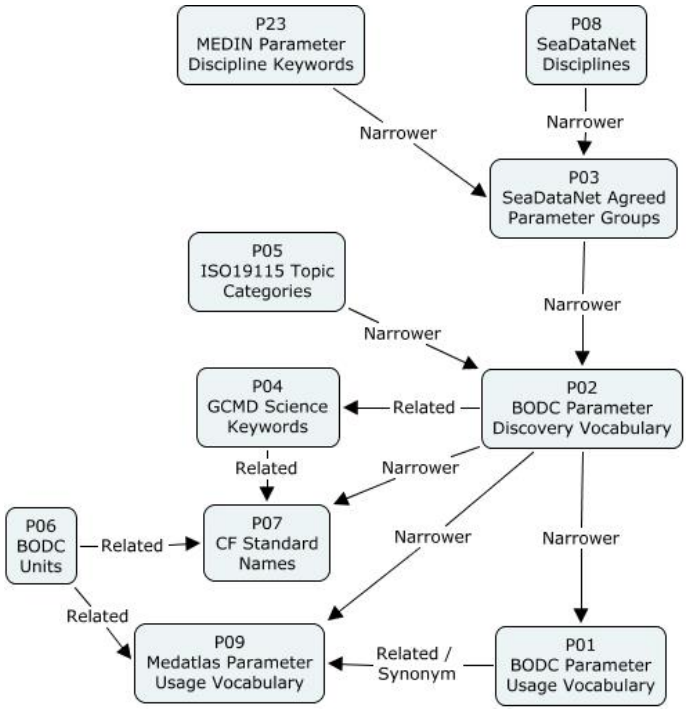
However, when building a hierarchical thesaurus, it is important that concepts defined at the same level of the hierarchy maintain a similar degree of granularity. If the thesaurus is imagined as a pyramid, making a concept at a given level too narrow or broad in its definition is like placing a too 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 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.

[6] 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.

[7] 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.



*An example from the NERC Vocabulary Server**Error! Bookmark not defined.** 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[8] netCDF metadata conventions group; the SeaDataNet and MarineXML Vocabulary Content Governance Group (SeaVoX)[9]; and the NETMAR ontology governance body[10]. The role of the content governance group is analogous to the International Organization for Standardization (ISO) definition of a "control body"[11].

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,

[8] http://cf-pcmdi.llnl.gov/

[9] https://www.bodc.ac.uk/data/codes_and_formats/seavox/

[10] http://netmar.nersc.no/

[11] http://www.dgiwg.org/Terminology/faq-other.php

XML); and the various versions of the concepts, collections and scheme are maintained and accessible. For example, within the SeaDataNet project this technical governance is provided by the British Oceanographic Data Centre as the developer and maintainer of the NERC Vocabulary Server**Error! Bookmark not defined.** (NVS). The role of the technical governance group is analogous to the ISO definition of a “register manager”[11].

Making the content available

The NERC Vocabulary Server

The NERC Vocabulary Server provides access to lists of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic and wider community.

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 controlled vocabularies delivered by the NERC Vocabulary Server contain the following information for each term:

- Key — a compact permanent identifier for the term, designed for computer storage rather than human readability
- Term — the text string representing the term in human-readable form
- Abbreviation — a concise text string representing the term in human-readable form where space is limited
- Definition — a full description of what is meant by the term

All of the vocabularies are fully versioned and a permanent record is kept of all changes made.

Connectivity

Consumers may access the Vocabulary Server either using the ReSTful URIs described below or via SOAP.

SOAP is a design of Application Programming Interface (API) for exchanging structured information across computer networks as the result of calls to web services. It relies upon XML (eXstensible Markup Language) documents for passing messages.

SOAP consumers should generate their client implementation from the Web Service Description Language (WSDL) documentation available athttp://vocab.nerc.ac.uk/vocab2.wsdl

SPARQL is standard query language for interrogating knowledge stores such as NVS2.0. The SPARQL endpoint may be found athttp://vocab.nerc.ac.uk/sparql from where queries may be entered directly and the return format chosen. Once users are comfortable with this interface and with building SPARQL queries, they may take the resulting URLs and use them to access the SPARQL endpoint programmatically.

Collection, concept and scheme URIs

Collections, concepts and schemes are presented to the Server as Uniform Resource Identifiers (URIs), or in this case actually URLs, in the following syntax.

Collections — A concept collection is useful where a group of concepts shares something in common, and it is convenient to group them under a common label. In NVS2.0, concept collections are synonymous with controlled vocabularies or code lists.

http://vocab.nerc.ac.uk/collection/
http://vocab.nerc.ac.uk/collection/colRef/colVer/
e.g. http://vocab.nerc.ac.uk/collection/P03/current/
http://vocab.nerc.ac.uk/collection/colRef/colVer/status/
e.g. http://vocab.nerc.ac.uk/collection/P03/current/accepted/

Concepts — A Simple Knowledge Organization System (SKOS) concept can be viewed as an idea or notion; a unit of thought. The notion of a SKOS concept is useful when describing the conceptual or intellectual structure of a knowledge organization system and when referring to specific ideas or meanings established within that system.

http://vocab.nerc.ac.uk/collection/colRef/colVer/conRef/
e.g. http://vocab.nerc.ac.uk/collection/P03/current/D005/

Schemes — A concept scheme can be viewed as an aggregation of one or more SKOS concepts. Semantic relationships (links) between those concepts may also be viewed as part of a concept scheme. A concept scheme is therefore useful for containing the concepts registered in multiple concept collections that relate to each other as a single semantic unit, such as a thesaurus.

http://vocab.nerc.ac.uk/scheme/
http://vocab.nerc.ac.uk/scheme/schemeRef/
e.g. http://vocab.nerc.ac.uk/scheme/ICANDIS/

where

- http://vocab.nerc.ac.uk/collection/ and http://vocab.nerc.ac.uk/scheme/ respectively provide catalogues of the available concept collections and concept schemes.
- colRef is an internal opaque identifier for the concept collection, e.g. P02 for the SeaDataNet Parameter Discovery Vocabulary.
- colVer may be a valid concept collection version number or 'current' to specify the latest version of the collection.
- status may be 'all', 'accepted' or 'deprecated' to indicate whether all concepts related to a collection should be returned, or only the accepted or deprecated concepts.
- conRef is an internal opaque identifier for the concept within the concept collection, e.g. TEMP for 'Temperature of the water column' in the SeaDataNet Parameter Discovery Vocabulary.
- schemeRef is an internal opaque identifier for the concept scheme, e.g. ICANCOERO for the International Coastal Atlas Network Coastal Erosion Thesaurus.

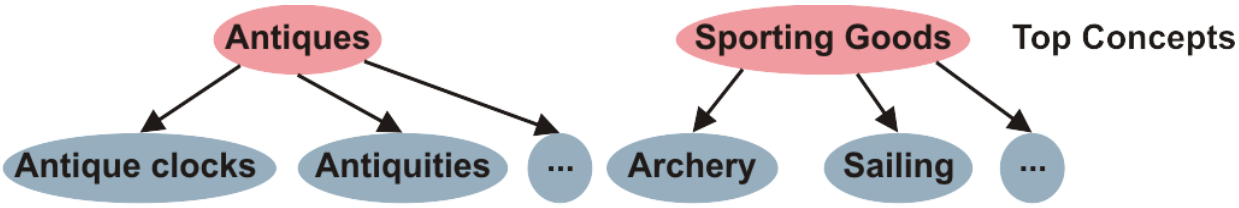
Simple Knowledge Organization System

On the NERC Vocabulary Server (NVS), knowledge organization systems are built upon the World Wide Web Consortium's Simple Knowledge Organization System[12] (SKOS) standard. SKOS is designed to provide a method for the online publication of controlled vocabularies and thesauri. The NVS platform is used by many projects and groups to publish collections of vocabulary terms and thesauri. 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

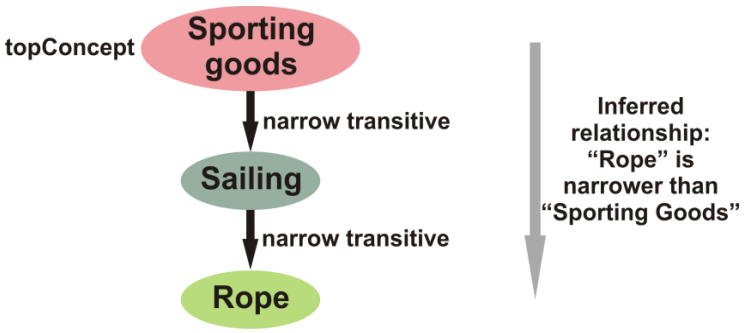
[12] http://www.w3.org/2004/02/skos/

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.

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 47). 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. Schemes are used on NVS as a discovery tool for concepts, and collections to store and publish concepts and for referencing their identifiers.

NVS has additionally extended the SKOS model to allow synonyms to be identified using the Web Ontology Language's[13] `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.

Adding Content to the NERC Vocabulary Server

Incorporating a Knowledge Organization System

The simplest way 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.

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:

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

{"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 on the NERC Vocabulary Server and further updates can be made by authorised persons through a web interface accessed from the British Oceanographic Data Centre website[14].

[14] https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_editor/

Accessing the Knowledge Organization System

Once deployed within the NERC Vocabulary Server, a knowledge organization system can be accessed in much the same way as a web site, using Uniform Resource Locators[15] (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/

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

[15] http://en.wikipedia.org/wiki/Url

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)[16] based access described in the associated Web Services Description Language (WSDL) document[17].

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**Error! Bookmark not defined.**; the European Environment Agency General Multilingual Environmental Thesaurus**Error! Bookmark not defined.**; and GeoNames[18]. Relationships between a concept in the NVS and any external concept can be specified in the same way as the internal mappings (see page 48) but with the NVS URL replaced by the URL of the external concept as the object of the relationship. For example:

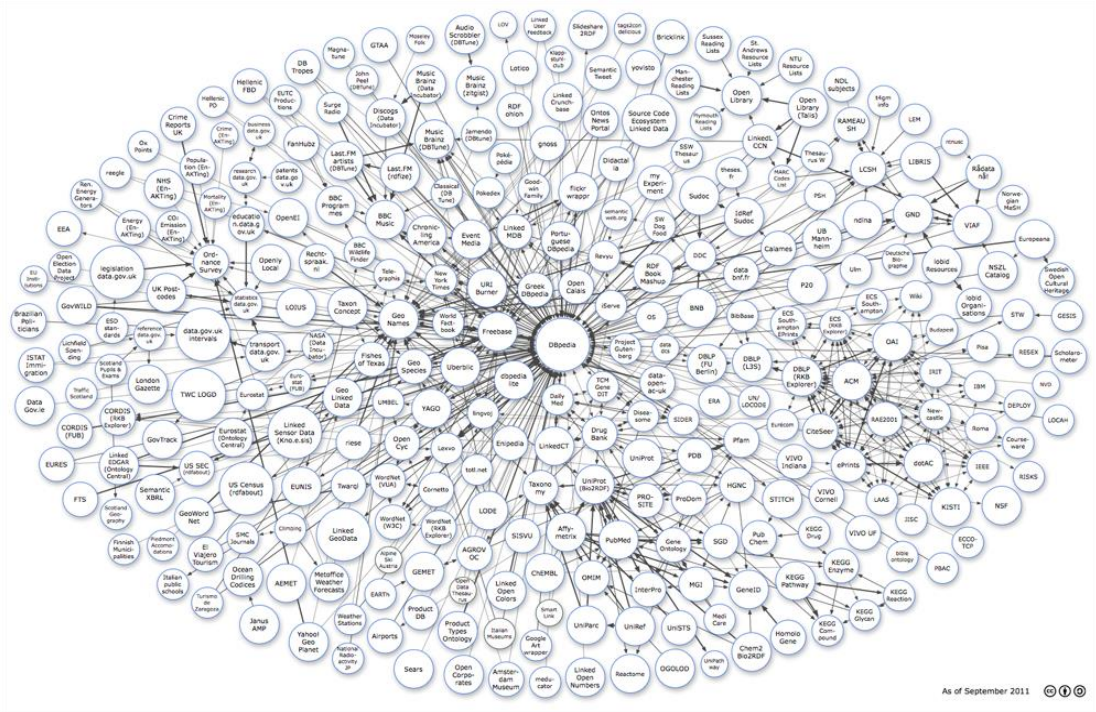
```
http://vocab.nerc.ac.uk/collection/P21/current/MS10360/ (sulphides)
"broader"
http://www.eionet.europa.eu/gemet/concept/4350 (inorganic substances)

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

[16] http://en.wikipedia.org/wiki/SOAP

[17] http://vocab.nerc.ac.uk/v2.wsdl

[18] http://www.geonames.org/



The Linking Open Data project cloud[19].

Incorporating knowledge organization systems in metadata

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.

Using CF-Standard Names to SeaDataNet Parameter Discovery Term links

Queries may be issued against the NERC Vocabulary Server using the SPARQL standard at http://vocab.nerc.ac.uk/sparql. Below is a template query which returns the SeaDataNet Parameter Discovery Terms related to the CF Standard Name "air_temperature".

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT ?standardNameLabel ?p02Uri
WHERE
{
   BIND ('air_temperature' as ?standardName)
   ?standardNameUri skos:prefLabel ?standardNameLabel
```

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

```
    FILTER CONTAINS(str(?standardNameUri),'/collection/P07/')
    FILTER (LANG(?standardNameLabel) = 'en')
    FILTER (str(?standardNameLabel) = str(lcase(?standardName))).
    ?standardNameUri ?rel ?p02Uri
    FILTER CONTAINS(str(?p02Uri),'/collection/P02/')
}
```

The URL of this query is
http://vocab.nerc.ac.uk/sparql/sparql?query=PREFIX+skos%3A%3Chttp%3A%2F%2Fwww.w3.org%2F2004%2F02%2Fskos%2Fcore%23%3E%0D%0ASELECT+DISTINCT+%3FstandardNameLabel+%3Fp02Uri%0D%0AWHERE%0D%0A%7B%0D%0A++++BIND+%28%27air_temperature%27+as+%3FstandardName%29%0D%0A++++%3FstandardNameUri+skos%3AprefLabel+%3FstandardNameLabel%0D%0A++++FILTER+CONTAINS%28str%28%3FstandardNameUri%29%2C%27%2Fcollection%2FP07%2F%27%29%0D%0A++++FILTER+%28LANG%28%3FstandardNameLabel%29+%3D+%27en%27%29%0D%0A++++FILTER+%28str%28%3FstandardNameLabel%29+%3D+str%28lcase%28%3FstandardName%29%29%29.%0D%0A++++%3FstandardNameUri+%3Frel+%3Fp02Uri%0D%0A++++FILTER+CONTAINS%28str%28%3Fp02Uri%29%2C%27%2Fcollection%2FP02%2F%27%29%0D%0A%7D&output=text The final &output parameter may be changed from “text” to “xml”, “json”, “csv” or “tsv” depending on which format the consumer wishes to receive.

As of 19th December 2014, 2123 CF Standard names are mapped to the SeaDataNet Parameter Discover Vocabulary, leaving 664 not mapped. The unmapped CF Standard Names are listed in MS9 document.