

# Exploiting and Extending Vocabularies for Faceted Browse in Earth System Science

Ruth E. Petrie<sup>a</sup>, Bryan Lawrence<sup>b,c</sup>, Martin Jukes<sup>a</sup>, Victoria Bennett<sup>a,d</sup>, Philip Kershaw<sup>d</sup>, Ag Stephens<sup>a</sup>, Alison Waterfall<sup>d</sup>, Antony Wilson<sup>e</sup>

<sup>a</sup>National Centre for Atmospheric Science, CEDA, Science and Technology Facilities Council, UK

<sup>b</sup>National Centre for Atmospheric Science, Department of Meteorology, University of Reading, UK

<sup>c</sup>Department of Computer Science, University of Reading, UK

<sup>d</sup>National Centre for Earth Observation, CEDA, Science and Technology Facilities Council, UK

<sup>e</sup>Scientific Computing Department, Science and Technology Facilities Council, UK

---

## Abstract

The earth system grid federation (ESGF) deployed a faceted browsing system for finding data within a large (petascale) globally distributed climate data archive. This system relied on a “Data Reference Syntax” which was designed initially for providing meaningful navigable identifiers for data from the fifth climate model intercomparison project (CMIP5). In this paper we provide additional context and expand on the need for such systems, discuss the inherent dependencies on controlled vocabularies, and show how the DRS concept has been extended to provide support for additional projects (CORDEX, CLIPC, CMIP6). We discuss the nature of search and browse and the impact of linked data concepts on successful data discovery. We also discuss the wider applicability in environmental science of faceted browse coupled with meaningful data identifiers.

**Keywords:** Earth System Science, Data, Metadata, Data Reference Syntax, Faceted searching

---

## 1. Introduction

Earth System Science (ESS) has always had difficulties associated with both the number and volume of data products. These difficulties arise because both earth observation and numerical simulation involve continually increasing data production driven by underlying trends in computing, always at the edge of what is possible, and so managing and manipulating such data at scale has always required heroic effort. For example the European Space Agency Sentinel satellites (Berger et al., 2012) are currently (2018) approaching 15 TB/day (5PB/year — Knowelden, 2019) and next iteration (phase 6) of Coupled Model Intercomparison Project (CMIP6) is anticipated to produce around 10-20 PB of data in the 2019-2020 timeframe.

One of the many challenges with this data is facilitating the discovery and use of climate data records. Climate data comes from a variety of sources such as modelling (including global, regional and seasonal modelling), earth observation (e.g. from satellites, in-situ observations, radiosondes etc.), reanalysis (a combination of both model and observational data) and climate impact indicators (derived products). All these different climate data products are produced within different parts of the ESS community and it is common for the different communities to have different data conventions (such as file formats) and standards. Nonetheless, these heterogeneous data need to be processed automatically, a problem which grows larger with increasing

volume and heterogeneity. Such automated processing requires the use of data description and organisation methodologies which are well defined and structured, which support discovery via multiple methods and entry points, and can be applied consistently across the ESS community.

In this paper we discuss two key methodologies: controlled vocabularies and structured syntaxes to support faceted browse, and their application in both the Climate Information Platform for Copernicus (CLIPC) and the Earth System Grid Federation.

Controlled vocabularies are sets of terms and definitions which have some managed process for change and maintenance which attempts to maximise accuracy, minimise ambiguity and repetition, and support incremental updates. They are typically managed in computer systems which provide both human and machine readable interfaces (e.g. the NERC vocabulary service available at [https://www.bodc.ac.uk/resources/products/web\\_services/vocab/](https://www.bodc.ac.uk/resources/products/web_services/vocab/), Latham et al. 2009; Leadbetter et al. 2012).

Faceted Browse is a particular form of data navigation which will be familiar to many people from application in internet shopping where a set of search results can be progressively refined by a set of pre-defined characteristics of the products (e.g. when shopping for fridges, refining by dimensions, manufacturer, energy rating etc). An important characteristic of faceted browse is that it is not hierarchical (e.g. for the fridge purchase example, one could start with manufacturer and refine, or start with energy rating and refine, etc).

*TODO: Ruth: Add paragraph on paper layout here, based on the following original text and the eventual structure of the paper* In this paper we are interested in how the known technologies of faceted search is applied and utilised in the ESS and how this could be applied to other scientific disciplines.

In order to achieve this two things are required. Firstly, a framework of both hardware and software infrastructure that supports for faceted searching is required and secondly a well defined set of controlled vocabularies that are widely adopted by the community.

It is essential that both the data and metadata are easily accessible to all users. In the context of big data this means not only that users are able to discover the information that they need but also that they are able to work with the data within their information handling systems. This consists of two key aspects: (1) Technical and domain specific terms must have accessible definitions and (2) Such definitions must be provided in a form that other peoples' software can work with.

The aims of this paper are to describe the issues around data discovery through faceted search in the climate modelling community through a demonstration of how the research project the Climate Information Platform for Copernicus (CLIPC) ?REF? utilised and extended the existing infrastructure to provide a single point of access to a variety of heterogeneous data.

### *1.1. Data Discovery*

The notion of data discovery is understood by different communities differently, and even within one community the concept can mean different things depending on application. For example, for some earth observation users, discovering a dataset would mean finding the set of all images from all sensors which are relevant for their location at a specific time, and for others, it might be to discover the set of all images from a particular sensor. How this can be supported is a function of the available metadata, the way the data is organised, and the software which provides the discovery service. In particular, the success or failure of such discovery is often dependent on how the data publisher has organised data into datasets, what navigation facilities are provided to find datasets, and whether-or-not and how, datasets can be subsetting.

There are two key steps involved in finding data which can be characterised as teleporting and orienteering (Teevan et al., 2004): the former involves "jumping" to the neighbourhood where

the right data can be found, and the latter, to the process of navigating around to find exactly what is wanted. These two steps need to be supported by data metadata. In practice there are several sorts of relevant metadata, particularly where data is kept in files on a disk or tape system as is often the case for high volume ESS data: metadata held in the files, metadata which appears in the physical layout of the files (e.g. directory names on a file system) and metadata held in databases or web-pages. Whatever their source and location, they can be characterised using the taxonomy introduced in Lawrence et al. (2009): “A for archive” data is necessary to navigate (orienteer) around a file system, potentially utilising information held in the files and the filepaths, “B for Browse” metadata is also used for navigation, but also to discriminate between similar datasets (e.g. two model simulations of the same phenomenon). “C-Character” holds information about actual or perceived quality, and “D-Discovery” provides the equivalent of dataset catalogue records, providing the information necessary for teleporting (to find a point from where one can orienteer).

If the datasets are not organised with sufficient granularity, or datasets structures differ, then many discovery use cases cannot exploit orienteering, either because the datasets are too large (one does not get close enough with teleporting for orienteering to work), or because the method of orienteering is too different between datasets. When the information system and/or data are predominantly hierarchical (organised in simple tree structures) this problem is exacerbated. While teleporting can arrive at some point “in the tree”, it becomes difficult to find similar parts of multiple trees via orienteering, unless the trees have similar structures. When the datasets have too much granularity, there are too many potential teleporting targets and so the metadata system needs to compensate so that it can produce aggregated views that can be unfolded into the constituent parts as orienteering proceeds. The obvious difficulty with aggregation is that there may be many different ways to do the aggregation.

Faceted browse with aggregation provides a sophisticated method of dealing with some of these issues, but only if it is underpinned by controlled vocabularies which ensure that different datasets can be viewed and/or aggregated using commonly understood terms. Datasets then need to be tagged with the correct combination of terms, and the information system (hardware and software) need to deliver the requisite functionality. Here we concentrate on the vocabularies and dataset granularities needed to deliver ...

*TODO: Ruth: COMPLETE.*

## **2. Data Reference Syntax**

### *2.1. History*

At the advent of the fifth climate model intercomparison project, CMIP5, it was apparent that the community was going to be faced with at least a petabyte of data organised into at least a million files. Data was being produced by multiple organisations, using different software systems (models) according to the needs of a variety of experiments, and consisting of hundreds of different output variables being produced by simulations of ocean, atmosphere, land surface, etc. It was going to be housed in distributed federation of data nodes, and users were going to be expected to find and download only the data of interest to them.

Experience from the earlier CMIP exercises had led to the knowledge that structured metadata was crucial, so the notion of quality A-metadata was already present, and it was known that mixing different variables within files could be problematic (greater chance for errors in any given file, more likelihood that users downloading files would be downloading data they didn't need).

However, it was also known that different groups liked to organise their data differently. The solution which arose was the “Data Reference Syntax” (DRS) introduced in Taylor et al. (2012), which provided a human and machine readable structured identifier for what became known as “atomic datasets” (nearly indivisible granules of data). The DRS identifier utilised controlled vocabularies to provide CMIP5 both landing points for teleporting, and a nomenclature for a host of important routes to aggregating the atomic datasets. It was also completely agnostic about file organisation (it is often necessary to remind people that the D does not stand for “Directory”) allowing different groups to organise their data as they preferred.

The set of vocabularies which were chosen to construct a CMIP5 DRS identifier are listed in table 1. An individual identifier is constructed by using vocabulary members constructed by concatenating them all in specific sequence separated by a dot:

```
<project>.<product>.<institute>.<model>.<experiment>.<time.frequency>.  
<realm>.<cmor_table>.<ensemble>.<version>
```

for example

```
CMIP5.MPI-M.MPI-ESM-LR.amip.mon.land.Lmon.r5i1p1.v20120529
```

Facet	Definition
project	Fixed as CMIP5
product	The type of output produced by the model.
institute	The climate modelling centre(s) or University responsible for the model.
model	The specific name of the climate model used.
experiment	The valid CMIP5 experiment short identifier.
time frequency	The temporal frequency of the output data, e.g. “mon” for monthly data.
realm	The earth system realm of the data, e.g. “atmos” for the atmosphere.
CMOR table	A lookup table that relates the frequency of a variable and its realm.
ensemble member	The specific ensemble member of the model run of the form r<L>i<M>p<N>where, L M and N are integers and r is for realisation; i for initialisation and p is for physics.
version	This is an ESGF version to uniquely identify the dataset and version control the data, it is given the form vYYYYMMDD.

Table 1: Facet definitions for CMIP5

## 2.2. The use of the DRS to support faceted browse in the ESGF

The CMIP5 data were distributed by the Earth System Grid Federation (ESGF), an international collaboration built on a shared experience developing and deploying software infrastructure over the last two decades. It was initially designed to handle the CMIP5 project alone (Williams et al., 2011), but has since grown to encompass a number of other projects, all deploying variants of the original DRS. It now hosts in excess of 20 PB, and has been integral to recent assessments made by the Intergovernmental Panel on Climate Change (IPCC).

The ESGF is a global system with nodes distributed around the world and all continents represented. Nodes interoperate with each other using a peer-to-peer paradigm, and provide the same

set of standardised data search and access protocols. Some nodes host data replicated from other nodes, but most simply provide local data for remote discovery and download.

The ESGF supports multiple projects, and the ESGF data discovery service relies heavily on the DRS concept, even for non-CMIP projects. The DRS identifier is the unique identifier which allows the system to know which datasets are replicated between nodes, and provides facets that both supports faceted browse in the discovery user interface and allows downstream applications to construct faceted search interfaces to ESGF data. An example of the faceted browse interface appears in figure 1.

The screenshot displays the ESGF CMIP5 faceted search interface. At the top, there is a navigation bar with logos for ESGF, CEDA, NERC, Science & Technology Facilities Council, and is-enes. Below this, the WCRP CMIP5 logo is prominent. The interface includes a search bar with a 'Search' button and a 'Display' dropdown set to '10 results per page'. A sidebar on the left contains faceted search options for Project, Product, Institute, Model, Experiment, Experiment Family, Time Frequency, and Realm. The 'Realm' facet is expanded, showing a list of datasets. The main content area displays search results, including a table with columns for Project, Product, Institute, Model, Experiment, Experiment Family, Time Frequency, and Realm. The results show three entries, all for CMIP5 datasets from the Max Planck Institute for Meteorology (MPI-M).

Figure 1: Screenshot of the Earth System Grid Federation (ESGF) CMIP5 faceted search interface. This example shows a case where no facets have been select so all results are returned. The CMIP5 `realm` facet is expanded on the left hand side showing the potential to select datasets from the seven different options in the CMIP5 realm vocabulary.

### 2.3. Underlying DRS principles for the ESGF

The experience in ESGF showed that the DRS concept of conflating an identifier with a compound set of facets was very powerful, providing both machines and humans real utility. However, this utility does not arise from chance, and it depends on the right set of facets, which in turn depend on how they are structured. Facet terms should either be terms from a controlled vocabularies or be of a flexible structured form (e.g. sub-ranges within an axis, such as periods within a date range).

A controlled vocabulary (CV) is simply a list of terms with an associated precise definition that must be replicated in a precise form (including spelling, case and other characters). Controlled vocabularies are widely used to organise and annotate large volumes of data, and are often used in file naming conventions, or to populate internal file metadata. For use as an identifier and as facets they need to be un-ambiguous, fully partition the space of possible terms, and be distinct

(no dataset may fall outside the domain of the vocabulary, or be capable of annotation by more than one term in each vocabulary). For example, the CMIP5 facet `Realm` spans the full set of high level modelling domains expected within an earth system model, and can only take one of the permitted values: `atmos`, `ocean`, `land`, `landIce`, `seaIce`, `aerosol`, `atmosChem`, and `ocnBgchem` (where the latter covers “ocean bio-geochemistry” and the rest should be self evident). Ensuring that these requirements are met is why they need “control” — it must be difficult to inadvertently break these criteria, yet also allow new terms to be added as would be necessary if either the domain is expanded, or the facets need sub-division.

Facets which encompass flexible structures similarly need to span the range with no overlapping, but they too need to be controlled to an extent to ensure the facets are meaningful across data providers.

As an example, the CMIP5 facet “Ensemble member” must unambiguously identify all possible ensemble members within an ensemble (a set of simulations which vary somehow across a particular experimental configuration). However, while the extent of the domain of such ensemble members may not be known except by the data provider (and hence is not encoded), the structure of the possible domain is prescribed, in this case to a triplet of the form  $r < L > i < N > p < M >$  where  $L$ ,  $N$  and  $M$  are integers and  $r$ ,  $i$  and  $p$  indicate realisation, initialisation, and physics axes respectively, so that ensemble members can be identified in the space of all simulations carried out by a particular provider along those axes. So `r3i2p1` indicates the third realisation of the second kind of initialised simulation with the same physics as all other `rLiMp1` instances.

It should be noted that some structured facets need to exist for identification purposes, but not all structured facets are actually of use for faceted browse, as for example in this case, one is unlikely to look for all simulations which have the facet `r3i2p1`. By contrast a time-period structured facet may be of use in true faceted browse.

Within ESGF the DRS has two further constraints: the initial controlled vocabulary for the first facet is the set of all projects with data on ESGF, and the last is a version number of the form `vYYYYMMDD` for that atomic dataset within the ESGF (multiple versions may occur as a dataset is updated or superseded). The ESGF software expects that different projects will expose different DRS structures (and facets), but that all datasets within a project are identified with the same DRS faceted syntax.

#### 2.4. Provenance Principles for a DRS

The use of the DRS within ESGF is primarily to allow data users to navigate amongst the available data using the data provenance during the browse phase of data selection. The initial application in support of CMIP5 was organised around the CMIP protocols, but a more generic approach is necessary for wider applicability, particularly if the DRS is to apply to observations as well as simulation.

In the typical production of a dataset, there is a series of processes and operations applied, analyses conducted, and interim data results generated; that is, a complex scientific workflow is enacted before a scientific experiment or observation yields its final data output. These processes and interim data outputs, along with other related metadata, form the dataset provenance. Provenance, also known as lineage, is increasingly important for determining authenticity and quality, especially when comparing products within the growing volumes of public domain datasets. It is also an important part of determining if data is fit for the intended purpose(s).

Observations & Measurements (O&M; Cox, 2016) provides a standards based framework for describing the characteristics of an event making an observation — what was measured, the

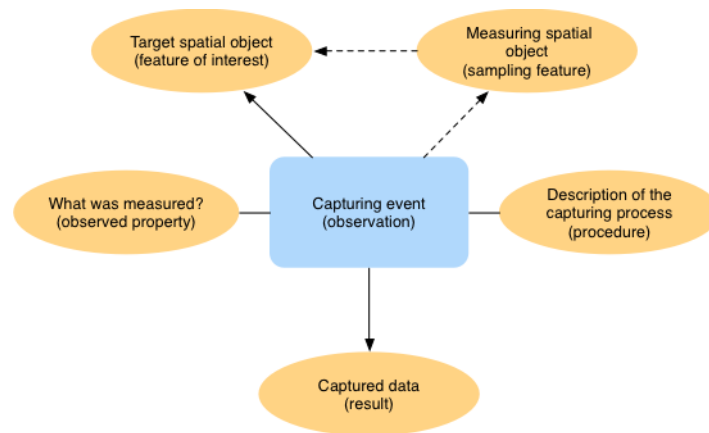


Figure 2: Observations and Measurements schema

procedure used, etc (see figure 2.4). As a framework it provides hooks for specialist descriptions which provide detailed descriptions of these key characteristics, but without prescribing so much that it becomes unimplementable. An integral concept within O&M is the concept of a “Sampling Feature” which explicitly allows for the spatiotemporal characteristics of the measurement to be recognised and captured.

Conforming with O&M provides a basis for extending DRS provenance concepts, by requiring that any DRS covers at least the following:

1. **Parties:** who is involved, ownership of the data; e.g. experimenter, institution;
2. **Procedure:** the process; e.g. model or instrument information;
3. **Sampling feature:** spatiotemporal information; e.g. the sampling frequency;
4. **Feature of Interest:** what is measured e.g. atmosphere, ocean, clouds;
5. **Observed property:** the parameter; e.g. air temperature.

These categories are not intended to be exclusive as there is not always a clean separation between them, particularly when one considers different perspectives. For example, in a modelling context cloud properties may be a feature of interest but in an observational context they may be the observed property.

The procedure used for an observation (or simulation) is crucial. One goal of any DRS will be to provide hooks for navigation from data to information about procedures used. Such navigation will depend on ancillary services which ideally share the same controlled vocabularies. For example, in the case of CMIP6 the facet “experiment” is shared by both the ESGF DRS and the Earth System Documentation system (<https://es-doc.org>), and it is possible to navigate between the DRS view of data in the ESGF and a description of the experimental protocol (Pascoe et al., 2019) at es-doc.org.

Provenance capture in a simulation workflow is relatively simple, however, the Advanced Climate Research Infrastructure for Data (ACRID, Shaon et al., 2012) project demonstrated that provenance capture is also possible for climate data observations (providing descriptions of data sources and versions, software versions, and processing options), but there is work to do to develop appropriate linking vocabularies.

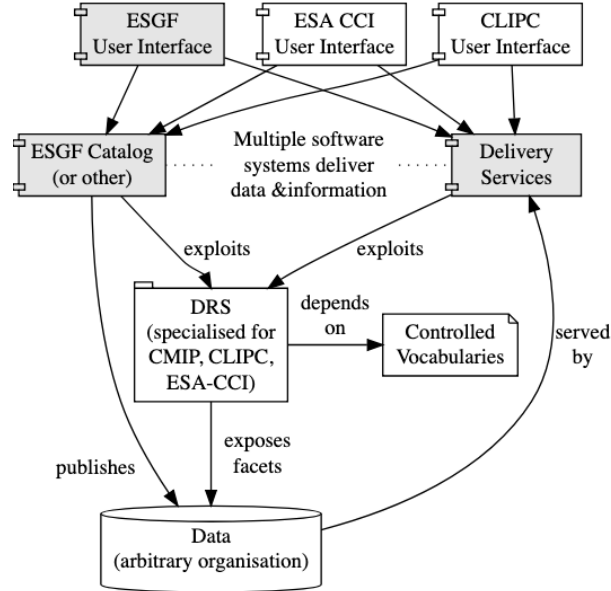


Figure 3: Exploiting the data reference syntax within a software system. Data is published into catalogs, and user interfaces (discussed in the text) exploit those catalogs and a range of delivery services to allow users to navigate around data offerings before selecting and acquiring data. Such systems can exploit a Data Reference Syntax and their controlled vocabularies.

### 3. Exploiting Data Reference Syntaxes in real systems

The CMIP5 DRS was deployed in ESGF, as described in section 2.1. The key components of the ESGF software and workflow are shown shaded in figure 1: a user interface exploits a catalog and delivery services. The DRS provides organising principles for the catalog, and may be exploited by delivery services (including ancillary information services as discussed above). In this section we discuss the extension of the original DRS to three important additional applications: to support the Climate Information Platform for Copernicus (CLIPC), the European Space Agency’s Climate Change Indicators portal, and the sixth CMIP phase, CMIP6.

#### 3.1. Climate information platform for Copernicus (CLIPC)

CLIPC includes a heterogeneous selection of data, extending the DRS considerably from the CMIP5 usage for global modelling by including selected datasets from the Regional Down-scaling Experiment (CORDEX), the Met Office Hadley Centre’s HadOBS project, and climate indicators calculated from the model data.

CORDEX was one of the first projects to be included in the ESGF as it expanded following CMIP5. It involved running regional climate models driven by boundary conditions from the CMIP5 archive for specific geographical domains (Giorgi et al., 2009).

The CORDEX DRS (described in Christensen et al., 2014) was a relatively straightforward extension of the CMIP5 DRS, with additional terms for domain, driving model, regional climate model, and regional climate model version — with the project facet constrained to be “CORDEX” (related projects with different DRS facets and project name also exist or will exist, including CORDEX-Adjust, Nikulin and Legutke 2016, and CORDEX2, Gutowski Jr. et al.



2016). For use within the CLIPC datasets, and associated DRS, not all the CORDEX terms were required as only one

<b>Facet</b>	<b>Definition</b>
<i>project</i>	Fixed as cordex.
<i>product</i>	The type of output produced by the model.
<i>institute</i>	The institute responsible for the data.
<i>domain</i>	A predefined region of the global that the data covers.
<i>driving model</i>	The specific name of the climate model used to provide the boundary conditions.
<i>experiment</i>	The valid CORDEX experiment short name.
<i>ensemble</i>	The ensemble member of the model run, inherited from the global model run; given in the form r<L>i<M>p<N>where, L M and N are integers and r is for realisation; i for initialisation and p is for physics.
<i>rcm_name</i>	The regional climate model name.
<i>rcm_version</i>	The regional climate model version.
<i>time frequency</i>	The temporal frequency of the output data.
<i>variable</i>	The short variable name identifier.
<i>version</i>	This is an ESGF version to uniquely identify the dataset and version control the data, it is given the form vYYYYMMDD.

Table 2: Facet definitions for CORDEX. Facets denoted with italics share the same controlled vocabulary as used for CMIP5, except for driving model, where that facet uses the CMIP5 source facet vocabulary. These facets are connected together using a “.” to construct the unique DRS: <project>.<product>.<domain>.<institute>.<driving\_model>.<experiment>.<ensemble>.<rcm\_name>.<rcm\_version>.<time\_frequency>.<variable>.<version>

A CMIP5 DRS example is

cordex.output.AFR-44.DMI.ECMWF-ERAINT.evaluation.r1i1p1.HIRHAM5.v2.day.uas.v20140804

The CORDEX data is hosted directly by the ESGF, with a subset indexed within CLIPC. The necessary DRS to work with the ESGF was established as a fairly direct extension from CMIP5 (table 2), and is subsumed directly into CLIP-C. The more widely used DRS elements that were required for the publication of model data were not always appropriate for observational data

In this section four different project DRS are considered; they are the CMIP5, Regional Downscaling Experiment (CORDEX), ESA Climate Change Initiative (CCI) and the Met Office Hadley Centre (MOHC) HadOBS projects. The CMIP5 project is a modelling project, CORDEX is a regional modelling project, the ESA-CCI project is an satellite observation project and the HadOBS project is a ground based observations project.

Table 3.3 shows the DRS elements used in each of the projects CMIP5, CORDEX, ESA-CCI and HadOBS respectively. The facets shown in normal font are unique facets for a given project the facets in grey italics are facets already defined and can be utilised by multiple projects. Since CMIP5 was the first project to use this approach to data discovery all the facets used were uniquely defined for this project.

These facets are connected together using a “.” to construct the unique DRS: <project>.<product>.<domain>.<institute>.<driving\_model>.<experiment>.<ensemble>.

<rcm\_name>.<rcm\_version>.<time\_frequency>.<variable>.<version> A CMIP5 DRS example is  
cordex.output.AFR-44.DMI.ECMWF-ERAINT.evaluation.r1i1p1.HIRHAM5.v2.day.uas.v20140804.

### 3.1.1. ESA-CCI: ESA Climate Change Initiative

The European Space Agency Climate Change Initiative (ESA-CCI) project is the first remotely sensed data to be published in ESGF and it required a number of new facets. Firstly consider the provenance facets. The term project is introduced and it is now common place to use project rather than activity and they are often used interchangeably. The product facet in the modelling community has a different meaning to that used in the EO community, therefore to work with the existing infrastructure the facet “product string” was included, where “product string” is the typical name of the EO product. The observation community also commonly have product versions associated with their data to keep up-to-date with the most recent observations and methodologies this was not required in the CMIP program and so an additional facet was introduced to describe this. There are also a number of additional procedural facets that are required to describe the ESA-CCI data, they are the processing level, sensor id and platform id.

Facet	Definition
project	Fixed as clipc
product	The type of output; esacci
cci project	The ESA CCI essential climate variable project.
time frequency	The temporal frequency of the output data.
processing level	The level of processing applied to the observational data, e.g. L3, L4.
CCI geophysical parameter	The observed quantity.
sensor	The instrument name.
platform	The satellite that carried the sensor.
product string	An additional product descriptor required for uniqueness, could be the name of a processing algorithm.
product_version realization	A version commonly associated with the dataset. The ensemble member.
version	This is an ESGF version to uniquely identify the dataset and version control the data, it is given the form vYYYYMMDD.

Table 3: Facet definitions for ESA CCI

The CCI facets were connected together to produce unique dataset identifiers of the form:  
`<project>.<cci_project>.<time_frequency>.<processing_level>.<cci_geophysical_parameter>.  
<sensor_id>.<platform_id>.<product_string>.<product_version>.<realization>.<esgf_version>`

A CMIP5 DRS example is  
`clipc.esacci.CLOUD.day.L3U.CLD_PRODUCTS.MODIS.Aqua.MODIS_AQUA.1-0.r1.v20120704.`

### 3.1.2. MOHC HadOBS: Met Office Hadley Centre, observational data products

The MOHC HadOBS observational data also required additional provenance information to describe the data effectively.

These facets are connected together using a “.” to construct the unique DRS:  
`<project>.<product>.<institute>.<framework>.<collection>.<frequency>.  
<realization>.<product_version>.<esgf_version>` A CMIP5 DRS example is  
`clipc.insitu.MOHC.HadOBS.HadISDH.mon.r1.v2-1-0-2015p.v20151231.`

Facet	Definition
project	Fixed as CLIPC
product	The type of data
institute	The institute responsible for the data.
framework	Dataset framework
collection	The dataset collection
frequency	The temporal frequency of the output data.
realization	The ensemble member.
product_version	A version commonly associated with the dataset.
version	This is an ESGF version to uniquely identify the dataset and version control the data, it is given the form vYYYYMMDD.

Table 4: Facet definitions for HadOBS

The screenshot displays the 'CLIPC at CEDA' search interface. At the top, there are logos for ESGF, CEDA, NERC, Science & Technology Facilities Council, and is-enes. Below the logos, a navigation bar includes 'Home', 'You are at the ESGF-INDEX1.CEDA.AC.UK node', and 'Technical Support'. The main search area features a sidebar on the left with facets: Project (clipc (2)), Product (insitu (2)), Institute (MOHC (2)), Framework (HadOBS (2)), Collection (HadISDH (2)), Domain, Driving Ensemble, Driving Reanalysis, Reanalysis, Reanalysis Ensemble, Time Frequency, Ensemble, Product Version, Variable, and Data Node. The search bar contains 'Enter Text:' and buttons for 'Search', 'Reset', 'Display 10 results per page', and 'More Search Options'. Below the search bar, the search constraints are listed: 'insitu | clipc | HadISDH | HadOBS | MOHC'. The results section shows 'Total Number of Results: 2' and a list of two datasets. Each dataset entry includes the dataset name, data node, version, total number of files, and links for 'Show Metadata', 'List Files', 'THREDDS Catalog', 'WGET Script', and 'Globus Download'.

Figure 4: Example ESGF search for insitu observational data

### 3.2. CMIP6

Although not a part of the CLIPC project, since that project took place the phase 6 CMIP data (CMIP6) is now being released and the DRS have been defined as shown in the table 3.2. Many similar facets are common between CMIP5 and CMIP6 however some have been renamed \*\* this is not good practice why oh why \*\*. The largest change from CMIP5 to CMIP6 is the granularity level of the DRS. In CMIP5 all variables for a given dataset were included within the dataset. There were often many 20-30 variables within a dataset. This was not optimal from a search or data management perspective. Therefore within CMIP6 the variable has been elevated to be a distinct facet. This is extremely useful given the much larger volume of CMIP6. For

example a user may simply be interested in a couple of variables for example sea ice and sea surface temperature, a user can simply search for these two variables and then narrow down their search from there.

Facet	Definition
mip_era	The phase of CMIP in this case CMIP6; equivalent to the CMIP5 project.
activity_drs	The model intercomparison project (MIP) to which the data belong.
institution_id	The climate modelling centre(s) or University responsible for the model.
source_id	The specific name of the climate model used.
experiment_id	The valid CMIP6 experiment short identifier.
member_id	The specific ensemble member of the model run of the form r<L>i<N>p<M>f<R>where, L, M, N and R integers and r is for realisation; i for initialisation; p is for physics and f is for forcing.
table_id	A lookup table that relates the frequency of a variable and its realm.
variable_id	A short variable name identifier.
grid_label	A short grid type identifier.
Version	This is an ESGF version to uniquely identify the dataset and version control the data, it is given the form vYYYYMMDD.

Table 5: Facet definitions for CMIP6

These facets are connected together using a “.” to construct the unique DRS: <mip\_era>.<activity\_drs>.<institution\_id>.<source\_id>.<experiment\_id>.<member\_id>.<table\_id>.<variable\_id>.<grid\_label>.<version>

A CMIP6 DRS example is

CMIP6.CMIP.AWI.AWI-CM-1-1-MR.historical.r5i1p1f1.3hr.rldscs.gn.v20181218.

### 3.3. Analysis of DRS extensions

## 4. Improving navigability

The benefit of using controlled vocabularies include flexibility, scalability and the linking of information subsystems.

One example of a mature controlled vocabulary within the Climate Science community is the Climate and Forecast (CF) standard names ?REF?. This is a list of variables names used in the climate and forecast community. Each term has precise spelling and definition. For example, air\_pressure\_at\_sea\_level has the definition “sea\_level means mean sea level, which is close to the geoid in sea areas. It is defined as having canonical units of Pascals (Pa). Having precise definitions means that meteorologists and climate scientists anywhere in the world using this standard name know that they are referring to the same quantity. This becomes ever more important when considering more complex variables for example radiative fluxes which have a vector component of direction and can be absolute or net. Having a name which clearly specifies the direction of the radiation and whether it is the absolute or net value is vital to ensure that variables are compared correctly or radiative budgets calculated correctly.

Using CVs is an essential component of a DRS, however CVs must be managed and the complexity of these can vary. In the simple example of the “realm” facet of CMIP5 there are only

	Provenance	Procedure	Sampling	Feature	Parameter
<b>CMIP5</b>	activity product institute	model experiment ensemble	frequency	realm cmor_table	variable name
<b>CORDEX</b>	<i>activity product institute</i>	domain driving_model experiment <i>ensemble</i> <i>model</i> rcm_version	<i>frequency</i>		
<b>HadOBS</b>	framework collection				
<b>ESA-CCI</b>	project product_string product_version	processing level sensor id platform id realization	time_frequency	cci_project	cci_geo_quantity*
<b>CMIP6</b>	mip_era activity model_cohort product institution_id	source_id experiment_id source_type variant_label	nom_resolution* sub-experiment grid label frequency	realm table id	variable cf_std_name*

Table 6: DRS elements classified by schema element. (terms with \* suffix have been abbreviated for presentation)

seven terms in the CV. Where there are only a small number of terms they could be managed for example in GitHub like many of the controlled vocabularies for CMIP6 as they require minimal management. In contrast the CF standard name table (a CV) currently consists of around 6000 terms and is managed by community experts. In order to add a new standard name to the table, it must be proposed, moderated (by the community experts) and approved; this involves a substantial amount of effort and collaboration. It is recommended that when using a CV in a DRS where possible the terms should be taken from existing CVs. In comparing the terms used in the “frequency” facet it has been noticed that different data producers sometimes use subtly different terminology. For example, it is not uncommon to see year and yr, or monthly and mon. While it is possible to use semantic web technologies such as SKOS to relate these terms it is most beneficial if terms were used consistently.

A number of new controlled vocabularies have been defined for CLIPC. The content of these new controlled vocabularies has been defined in consultation with the data providers and curators. Provenance information is currently represented using PROV-O for new vocabularies to be incorporated into the NERC Vocabulary Server (NVS). The new vocabularies for CLIPC included defining conceptual schemes or themes such as the Global Climate Observing System (GCOS) Essential Climate Variable (ECV) domains and subdomains: atmospheric, terrestrial, atmospheric surface, atmospheric upper-air, atmospheric composition, oceanic surface, oceanic sub-surface. To reconcile the different vocabularies for the different climate data records the Simple Knowledge Organisation System (SKOS) is used to provide a mapping framework that links terms from different vocabularies using semantic mappings.

SKOS provides relational matches between two predefined vocabularies using the Resource Description Framework (RDF). The SKOS relationships between different vocabularies can be loosely defined as

- associative: concepts are related, they may be approximately interchangeable; can be either close or exact relationships
- hierarchical: concepts can be broader or narrower:
  - broader: the current term has a more specific definition than the related term e.g. carbon dioxide has a broader relationship to greenhouse gases
  - narrower: the current term has a less specific definition than the related term e.g. atmospheric composition has a narrower relationship to greenhouse gases

Using SKOS the internal hierarchical and associative mappings are defined, this allows terms within and across different controlled vocabularies to be related greatly enriching the data search experience.

#### *4.1. The Climate Change Initiative (CCI) example*

The data reference syntax that were defined in 3.1.1 for the ESA CCI project were used not only in the CLIPC project but also in the ESA CCI portal. Here a faceted search was implemented on the data utilising the linked data technologies to enhance search and discovery. Example:

Screenshot:

## **5. Summary and Future Work**

The first use of DRS in the CMIP5 project provided a comprehensive list of facets that were relatively simply managed. A DRS:

- provides a unique identifier for the dataset,
- provides a common terminology for a collection of datasets,
- can aid filesystem management,
- allows faceted searching.

Since then many new projects have emerged that use the ESGF infrastructure for publication and thus need a good DRS in order to provide a good quality faceted data search. A number of problems are now emerging and they are detailed below.

There must be robust communication in this multi-disciplinary environment as this is a community exercise with technical constraints.

#### *5.1. Lessons learned*

#### *5.2. Governance*

- The importance of maintenance of information . . . facet name inconsistencies eg table, table\_id - Difficulties of a globally distributed differential funded environment on information management. Having all the controlled vocabularies for each project stored in a central GitHub repository - even this has its problems... - social concept of communications costs necessary ... link to the vocabularies as being key to this.

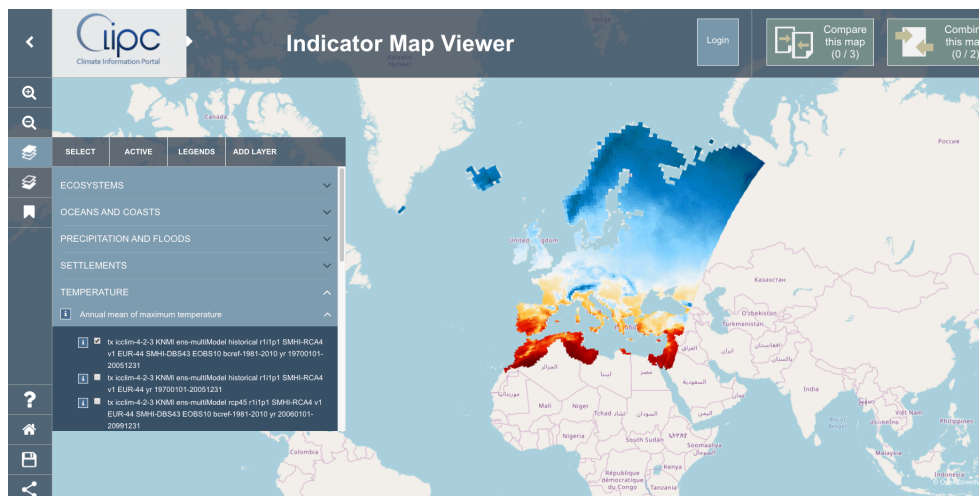


Figure 5: Mapping tool from CLIPC toolbox (If we use this one will need to introduce the DRS for the impact indicators)

### 5.3. Flexibility and Futures

#### SOME FIGURES

The CLIPC portal allows users access to a visualisation tool


Main CLIPC SEARCH

CLIPC as found in C4I

CCI

#### References

- Berger, M., Moreno, J., Johannessen, J.A., Levelt, P.F., Hanssen, R.F., 2012. ESA's sentinel missions in support of Earth system science. *Remote Sensing of Environment* 120, 84–90. doi:10.1016/j.rse.2011.07.023.
- Christensen, O., Gutowski, W., Nikulin, G., Legutke, S., 2014. CORDEX Archive Design. Technical Report. Danish Meteorological Institute.
- Cox, S.J.D., 2016. Ontology for observations and sampling features, with alignments to existing models. *Semantic Web* 8, 453–470. doi:10.3233/SW-160214.
- Giorgi, F., Jones, C., Asrar, G.R., 2009. Addressing climate information needs at the regional level: The CORDEX framework. *WMO Bulletin* 58(3), 175–183.
- Gutowski Jr., W.J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., Tangang, F., 2016. WCRP COordinated Regional Downscaling EXperiment (CORDEX): A diagnostic MIP for CMIP6. *Geosci. Model Dev.* 9, 4087–4095. doi:10.5194/gmd-9-4087-2016.
- Knowelden, R., 2019. Sentinel Data Access - Annual Report 2018. Technical Report COPE-SERCO-RP-19-0389. SERCO.
- Latham, S., Cramer, R., Grant, M., Kershaw, P., Lawrence, B., 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: Mathematical, Physical and Engineering Sciences* 367, 1015–1019. doi:10.1098/rsta.2008.0238.
- Lawrence, B.N., Lowry, R., Miller, P., Snaith, H., Woolf, A., 2009. Information in environmental data grids. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367, 1003–1014. doi:10.1098/rsta.2008.0237.
- Leadbetter, A., Lowry, R., Clements, O., 2012. The NERC Vocabulary Server: Version 2.0, in: *Geophysical Research Abstracts*, EGU, Vienna, Austria. pp. EGU2012–2943.



Climate Information Portal

[Home](#) | [Getting started](#) | [Access climate data](#) | [Impact indicators](#) | [MyCLIPC data processing](#) | [Project information](#)

MyCLIPC

Login for more services!

CLIPC: Constructing Europe's Climate Information Portal

CLIPC provides access to Europe's climate data and information.

[Home](#) > [Search Climate datasets](#)

### Search climate datasets

Search by

Free search

Date [dd/mm/yyyy]

Geobox

Parameter Group

Alkalinity, acidity and pH of the water column (ALKY)  
Ammonium and ammonia concentration parameters in water bodies (AMON)  
Chlorophyll pigment concentrations in water bodies (CPWC)  
Concentration of suspended particulate material in the water column (TSED)  
Date and time (ATMD)  
Dissolved metal concentrations in the water column (MTWD)  
Dissolved oxygen parameters for sediments (DGPA)  
Dissolved oxygen parameters in the water column (DOXY)  
Dissolved total and organic nitrogen concentrations in the water column (TDNT)

Reset all Send button







Data source	Total	To data
 <b>CLIPC catalogue</b> All climate (indicator) datasets produced and processing by CLIPC experts	493	→
 <b>SeaDataNet dataproducts</b> Marine data products like marine climatologies created within the SeaDataNet project.	17	→
 <b>EMODnet</b> Marine chemical data products created within the EMODnet Chemistry project.	156	→
 <b>MyOcean</b> Marine data products and forecasts available via CMEMS - the marine component for Copernicus	143	→
 <b>ESGF records</b> Climate data accessible via the Earth System Grid Federation infrastructure.		→

Figure 6: CLIPC data search



is-enes  Exploring climate model data in | IS-ENES | Contact | Sign in

Home Data discovery Downscaling Documentation Help About us Sign in

Search Catalogs Explore your own catalogs or files Map & Plot Processing

Filters ? Help

Project (1) Parameter (44) Frequency (4) Domain (1) Access (4) Date Geobox Free text [show all filters](#)

[clear all filters](#)

Quick select Project All Project properties (1)

CMIP

☐ GCM-data [CMIP5-project](#)

☐ GCM-data [CMIP5-project](#)

☐ GCM-data [NEXGDDP](#)

CORDEX

☐ RCM-data [CORDEX](#)

☐ RCM-data [CORDEX-Adjust](#)

OBSERVATIONS

☐ satellite-data [obs4MIPs](#)

☒ station data [CLIPC project](#)

Selected filters

☒ Project : clipc

Found 114 datasets. Displaying page 1 of 5.

[Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) [Next](#) [Export to CSV](#)







	clipc.RegRean.EUR-05.SMHI.SMHI-HIRLAM.v1d1-v1d2.SMHI-MESAN.v1.mon.tasmin.v20150601
	clipc.RegRean.EUR-05.SMHI.SMHI-HIRLAM.v1d1-v1d2.SMHI-MESAN.v1.mon.tasmax.v20150601
	clipc.RegRean.EUR-05.SMHI.SMHI-HIRLAM.v1d1-v1d2.SMHI-MESAN.v1.mon.tas.v20150601
	clipc.RegRean.EUR-05.SMHI.SMHI-HIRLAM.v1d1-v1d2.SMHI-MESAN.v1.mon.pr.v20150601
	clipc.RegRean.EUR-05.SMHI.SMHI-HIRLAM.v1d1-v1d2.SMHI-MESAN.v1.day.tasmin.v20150601

Figure 7: CLIPC data search through Climate for impacts portal

→ CCI SEARCH

Climate data search interface for the ESA Climate Change Initiative



ECV (2)

aerosol (2)

Search text (optional)

Clear facets

Search

Institute	Rutherford Appleton Laboratory	×
ECV	aerosol	×

2 results

↑ Dataset Information

📄 Product Guide

📄 FTP Download

📅 01 Jun 1995

📅 04 Jun 2003

ESA Aerosol Climate Change Initiative (Aerosol CCI): Level 3 aerosol products from ATSR2 (ORAC algorithm), Version 3.02

The ESA Climate Change Initiative Aerosol project has produced a number of global aerosol Essential Climate Variable (ECV) products from a set of European satellite instruments with different characteristics. This dataset comprises the Level 3 aerosol products from ATSR-2, using the ORAC algorithm, version 3.02. Both daily and monthly gridded products are available For further details about these data products please see the linked documentation.

**Data were processed by the ESA CCI Aerosol project team and supplied to CEDA in the context of the ESA CCI Open Data Portal Project.**

Additional Download Options

1

esacci.AEROSOL.day.L3C.AER\_PRODUCTS.ATSR-2.ERS-2.ORAC.03-02.r1.v20170402

Total: 2570

[ Show Files ]

[ THREDDS Catalog ]

[ Download Script ]

2

esacci.AEROSOL.mon.L3C.AER\_PRODUCTS.ATSR-2.ERS-2.ORAC.03-02.r1.v20170402

Total: 89

[ Show Files ]

[ THREDDS Catalog ]

[ Download Script ]

↑ Dataset Information

📄 Product Guide

📄 FTP Download

📅 20 May 2002

📅 30 Apr 2012

ESA Aerosol Climate Change Initiative (Aerosol CCI): Level 3 aerosol products from AATSR (ORAC algorithm), Version 3.02

The ESA Climate Change Initiative Aerosol project has produced a number of global aerosol Essential Climate Variable (ECV) products from a set of European satellite instruments with different characteristics. This dataset comprises the Level 3 aerosol products from AATSR, using the ORAC algorithm, version 3.02. Both daily and monthly gridded products are available For further details about these data products please see the linked documentation.

**Data were processed by the ESA CCI Aerosol project team and supplied to CEDA in the context of the ESA CCI Open Data Portal Project.**

Figure 8: CCI

18

- Nikulin, G., Legutke, S., 2016. Data Reference Syntax (DRS) for Bias-Adjusted CORDEX Simulations. Technical Report. Swedish Meteorological and Hydrological Institute.
- Pascoe, C., Lawrence, B.N., Guilyardi, E., Jukes, M., Taylor, K.E., 2019. Designing and Documenting Experiments in CMIP6. *Geoscientific Model Development Discussions*, 1–27doi:10.5194/gmd-2019-98.
- Shaon, A., Callaghan, S., Lawrence, B., Matthews, B., Osborn, T., Harpham, C., Woolf, A., 2012. Opening Up Climate Research: A Linked Data Approach to Publishing Data Provenance. *International Journal of Digital Curation* 7, 163–173. doi:10.2218/ijdc.v7i1.223.
- Taylor, K.E., Balaji, V., Hankin, S., Jukes, M., Lawrence, B., Pascoe, S., 2012. CMIP5 data reference syntax (DRS) and controlled vocabularies.
- Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R., 2004. The perfect search engine is not enough: A study of orienteering behavior in directed search, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 415–422.
- Williams, D.N., Lawrence, B.N., Lautenschlager, M., Middleton, D., Balaji, V., 2011. The Earth System Grid Federation: Delivering globally accessible petascale data for CMIP5, in: *Proceedings of the 32nd Asia-Pacific Advanced Network Meeting*, New Delhi. pp. 121–130. doi:10.7125/APAN.32.15.