CLIPC MS 19: Extended controlled vocabularies

Table of Contents

Abstract

Vocabularies are used both in the organisation of data and in supporting structured user access. This milestone reviews vocabularies in use across different areas of climate science and analyses their differences. There are differences in content, but also differences in logical structures and governance processes.

Executive Summary

Well defined vocabularies play a central role in the creation of well structured archives and data services. The vocabularies define the terminology used to describe and organise the data. Scientific domains such as chemistry and biology have well established global recognised vocabularies, but there is still a considerable degree of variation in the vocabularies used in the Earth system sciences. The status quo is reviewed. There are a variety of governance mechanisms in play. The WMO GRIB codes, for instance, go through an extensive consultation ensuring all members of the WMO have a chance to comment and are able to meet implementation requirements.

1 Introduction

This milestone examines in more details some of the vocabularies associated with the data collections discussed in Deliverable 5.1 “Climate Dataset Inventory”.

The CLIPC Architecture Design Document will set out the technological implementation choices in detail, this document will discuss the structure of the information in the vocabularies, and the options for organising and representing that structure.

Well defined vocabularies play a central role in the creation of well structured archives and data services. The vocabularies define the terminology used to describe and organise the data. The objective of these vocabularies is to provide a set of terms which can be used to uniquely specify every dataset held in the archive. Some of the terms will be generic, others specific to a particular activity or group of activities.

Scientific domains such as chemistry and biology have well established global recognised vocabularies (e.g. IUPAC nomeclature for chemical names), but there is still a considerable degree of variation in the vocabularies used in the Earth system sciences. The vocabularies used by the climate modelling community are different in many respects from those used in the context of numerical weather prediction and Earth observation.

This document will review existing practice and discuss the steps that CLIPC can take to support users and create a clear pathway to exploration of data resources from a disparate range of scientific communities.

The CLIPC portal will provide access to data from climate models, space borne instruments, in situ observations and re-analyses. While these different categories of data all describe the same Earth system, the repositories that manage the data reflect the usages and priorities of the different specialities.

Some differences are direct consequences of the different sources of data: e.g. data from satellites is associated with a particular platform and instrument. Other differences are more subtle, e.g. climate models produce hundreds of data variables, organised into “realms” – climate data from satellites is structured according to Essential Climate Variables (ECVs).

A list of vocabularies in use and under development is given in Appendix 1.

1.1 Issues

The vocabularies do not exist in isolation – part of the problem is to understand how they relate to each other, and how they are used to create, within each domain, a robust data referencing system. In many cases there is redundancy (in the form of multiple vocabularies covering similar concepts) to provide flexibility. Box 2 defines a “complete reference set” in terms of what is needed to organise the data and a “search” set as an extended set used to provide flexibility. This distinction between what is needed for robust data curation and what is desirable for flexible search will be discussed further below.

> ***Box 2: Vocabulary collections***
>
> **Complete reference vocabulary set**
>
> A reference vocabulary set is considered complete if specifying all the terms provides a unique specification of a variable.
>
> **Search vocabulary set**
>
> There are a variety of overlapping vocabularies in common use to define data: in order to provide a flexible data discovery service it is desirable to have a collection of search terms which is broader than those required for completeness. This may be done by including additional terms in the file metadata, by generating additional terms for inclusion in metadata records, or through an intelligent search interface.

The challenge for CLIPC is to produce a coherent framework spanning a broad range of activities.

1.2 Approach

The variety of standards described below creates an obstacle for users wanting to exploit a broad range of data products. It is clear that CLIPC cannot expect to replace standards which have broad international usage and extensive software frameworks built around them. The primary aim here is to document the vocabularies and collect them into a common searchable document – the CLIPC Book of Terms. In order to add value to this Book of Terms, cross-references between terms will be generated. The Book of Terms will thus not just be a document for users to browse, but may be integrated into the data discovery services (so that a user searching for a term which is in common usage in the climate modelling community can be directed towards data labelled with analogous terms from the Earth Observation community). This will be built on the NERC vocabulary service.

A critical aspect of the harmonisation of data will be the harmonisation of data formats and data format compliance checks. The Book of Terms must also be able to support compliance checking software. In fact, the use of the Book of Terms will provide an initial test that the implementation is actually usable by software applications.

> Facets: in informatics and in this document, facets refer to a set of vocabularies which can be used to organise and navigate collections of digital objects. Facets are widely used in commercial web sites to allow flexible navigation through an extensive range of products, replacing the older approach of organising products into a structured set of departments.

1.3 Outline

Section 2 will review the major existing vocabulary frameworks. In section 3, some frameworks for comparison and rationalisation are discussed. In section 3.4 a proposal for a sustainable system of managing links between independently maintained vocabularies is discussed.

There are a number of standards setting initiatives which underpin the implementation frameworks discussed in section 2: thee cross-cutting standards (e.g. NetCDF CF, GRIdded Binary – GRIB, ISO …, NetCDF standard global attributes, NASA, AGU, ….....) are reviewed in Appendix 2. The Simple Knowledge Organization System (SKOS - http://www.w3.org/2004/02/skos/) terms which will be used to relate vocabulary elements are also reviewed in appendix 2.

2 Existing Climate Data Frameworks

This section reviews vocabulary frameworks for publishing climate data: a framework is here considered to be a collection of vocabularies which together provide the means of structuring data archives and presenting users with structured access routes. These frameworks rely on a mixture of vocabularies which are specific to a specific collection of climate data and more generic vocabularies which may be subject to external governance processes. Appendix 3 reviews some of the major generic vocabularies.

2.1 Global Climate Models – the CMIP standards

The Coupled Model Inter-comparison Project (CMIP) has provided the core set of global climate projections which underpin our current understanding of global change. CMIP draws in contributions from major climate modelling centres around the world, specifying data standards so that the collected output can be analysed efficiently by a diverse community of research scientists.

The CMIP standards specify variable names, frequency of output, temporal averaging as appropriate, vertical coordinate information and a number of additional informative attributes for each variable. A software

package, CMOR, developed and maintained by PCMDI, is widely used to support creation of data files which are compliant with the standard.

2.2 Regional Climate Models – CORDEX

The CORDEX data follows the general structure of the CMIP standards, with the addition of vocabularies to define the driving models and the spatial domain of the regional climate simulations.

2.3 Global Re-Analysis products.

Global re-analysis products follow the conventions of the Numerical Weather Prediction centres within the WMO. The data is produced and distributed in the GRIB format. CLIPC contributed to a recent workshop organised by ECMWF to discuss the challenges of converting from GRIB to NetCDF. The main challenges are associated with the transformation of the metadata from one set of meta-data standards to another. The GRIB standards are in transition: the existing GRIB codes will be replaced by a more extensive and structured set of codes under GRIB2. See Appendix for more discussion of GRIB 1 and 2. CLIPC will support harmonisation of these standards through inclusion of the GRIB2 parameter codes in the CLIPC Book of Terms.

2.4 Regional Re-Analysis

The scientific and computational frameworks of regional re-analysis are closely related to those of regional climate modelling, so it is natural to base the approach to vocabularies on those defined for CMIP and extended for CORDEX.

Development of vocabularies for regional re-analysis is being led by SMHI as a joint activity between UERRA and CLIPC.

2.5 Obs4MIPS

The obs4mips project provides a route for the publication of selected observational datasets in a format which is compatible with the CMIP standards. Obs4MIPS uses the CMIP vocabulary framework (see discussion document by D. Clifford (in prep.)).

2.6 ESA Climate Change Initiative

The CCI standard is designed to provide unique file names of the form:

<Indicative Date>[<Indicative Time>]-ESACCI-<Processing Level>_<CCI Project>-<Data Type>-<Product String>[-<Additional Segregator>][-v<GDS version>]-fv<File version>.nc

or

ESACCI-<CCI Project>-<Processing Level>-<Data Type>-<Product String>[-<Additional Segregator>]-<IndicativeDate>[<Indicative Time>]-fv<File version>.nc

The terms in this filename template provide a complete reference set. Terms in the file metadata attributes provide additional ways to reference the data. Table 1 lists the filename terms, with some comments, and some additional terms from the file metadata. The ESA CCI standards have been informed by and make reference to earlier standards set by the Group for High Resolution Sea Surface Temperature (GHRSST) - See more at: www.ghrsst.org

Table 1. Complete reference vocabularies (**bold**) and search vocabularies

	Description	**Comments**

Programme		Fixed: ESA-CCI
CCI project	The project producing the data: corresponds to ECV	For a general purpose interface, it will be clearer to label this as the ECV.
Processing level	EO processing level, using standardised CCI definitions.	Need to check consistency with processing levels used by other EO datasets being considered for inclusion in CLIPC.
Data Type	Physical variable, e.g. Aerosol optical depth	There may be corresponding CMIP variables with different names – once a CF standard name has been selected for each data type it will be relatively easy to check for such correspondences. Once this is done, the CMIP names can also be added to the search record*.
Product	A string sometimes identifying the product in terms of the input dataset, sometimes referring to an algorithm	e.g. AATSR_ENVISAT(Data from AATSR on ENVISAT).
Additional Seg.	Structured text, following a complex set of rules. This text ensures that different post-processing choices (e.g. aggregation to monthly data) produce different file names.	
Indicative date		This information is not currently used in the ESGF search interface, but support for searches on a time range may be added in the near future.
Indicative time		
Data specification version	As used in GHRSST.	In the CMIP5 archive the data specification version is recorded in the file, but is not used in the discovery metadata.
File version	Need to clarify if reprocessing with a new data specification version should also result in an incremented file version number.	To avoid confusion, it may be that only the latest version is made visible through ESGF – but other versions may be retained to be available if needed to reproduce results.
Platform	Satellite(s) from which data have been included in the ECV	Specified in file metadata attributes
Sensor	Sensor(s) from which data have been included in the ECV	Specified in file metadata attributes
Algorithm	May be indicated in various places.	
Standard name	Specified in file.	
Frequency	Frequency is a key component of	For level 3 and 4 data the

	the data discovery options for climate model data.	time_coverage_resolution attribute carries the information. For level 2 data it is not clear that the concept makes sense: the time_coverage_resolution then records the frequency of observations along the track, not the frequency of measurements at a point. Proposal: leave it unset of level 2 and lower.
Spatial resolution	Following the approach set an indicative resolution in km.	Useful for users who may want to focus their search on a particular range of resolutions. Current ESGF interface does not provide support for range searches – but expect this within a few years. Use the metadata attributes: geospatial_lon_resolution and geospatial_lat_resolution.
Variable name in file	This is not specified in the CCI data specifications. It is specified in the file metadata attributes (variable name) but the standard does not define how the "main" parameter is distinguished from ancillary parameters in the file	In order to support automated processing (by users and archive managers) a table specifying the data variable names used will be generated.
Equivalent CMIP name	Where possible, a corresponding CMIP variable name.	

*Search record: when data is published, a "search record" is generated, based on file metadata. For robust publication procedures it is desirable that most of the content of the search record should be a simple reflection of the file metadata, but it is possible to add additional terms.

2.7 EUMETSAT CMSAF

The CMSAF data is published using a well developed set of vocabularies. It will be desirable to achieve some convergence with the ESA-CCI standards, but that cannot be dealt with in CLIPC. EUMETSAT uses 3 character codes for each variable – variables are listed in deliverable 5.1. As for the GRIB data, we will support the user community through inclusion of terms in the Book of Terms.

2.8 In situ observations

The range of datasets is more limited, and greater specialist attention is focussed on each parameter. The range of parameters for which global long term data sets are available is limited. CLIPC will develop a for a vocabulary framework to support publication of HadOBS data, exploiting experience gained from publication of ECV data from the ESA CCI programme.

2.9 Inter-Sectoral Impact Model Intercomparison Project (ISIMIP)

The ISIMIP (https://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/research/rd2-cross-cutting-activities/isi-mip) data has been published through ESGF (http://esg.pik-potsdam.de/esgf-web-fe/). Publication of the data relies on a number of vocabularies relating to forcing and impact.

2.10 Expert Team on Climate Change Detection and Indices (ETCCDI)

ETCCDI (http://www.clivar.org/panels-and-working-groups/etccdi/etccdi.php – a CLIVAR/Ccl/JCOMM/GEWEX expert team) compiled a list of Tier 1 indicators (statistics of the physical climate). 27 core indices have been defined (http://etccdi.pacificclimate.org/list_27_indices.shtml). At this point the terms defined have not been assembled into a complete framework, but it is included here as the existing vocabularies will be important for development of a framework within CLIPC.

2.11 CLIPC review of indicators

Deliverable 7.1 has created an initial review of indicators, and a structure for organising them. A key starting point is a split into three "tiers" which span the range from statistics of the physical climate to estimates of societal impact. Further work is needed on documentation of terminology.

3 Review

3.1 Maturity

Definitions

Vocabularies exist in some form in all branches of climate science, but there are various states of maturity in terms of consolidation into a complete set which can be consistently used to describe a wide range of data products. Here we look at 4 phases:

Maturity	Description
Complete conceptual framework	At this stage all the vocabularies are clearly defined.
Implementation plan	Once the vocabularies are defined it is necessary to create a set of conventions for file naming, data storage and metadata creation.
Deployment	Creation of an operational system is the ultimate test that the vocabularies are really complete and meet the user expectations. The latter is important as the vocabularies provide the route for data discovery, and if the terms that the users expect to see are not there it will cause inconvenience.
Linked	Deployed in a linked-data sense: terms are available through global standards, together with specification of appropriate links to other vocabularies.

Status summary

For climate model inter-comparison projects in which many different modelling teams provide independent realisations of exactly the same physical variables it is essential to have all details of the data format agreed before data production starts. In Earth observation, on the other hand, there is a significant tendency for scientists to specialise on a specific group of variables.

The SeaDataNet vocabularies have the best claim to "Linked" status, but links to widely used North American vocabularies are still not in place. The methodological approach followed by SeaDataNet, built on a SKOS vocabulary server, is nevertheless a clearly positive demonstration of how a truly linked system might work.

	Summary	Status	Role of CLIPC (and other projects)
Global climate models	Deployed for the CMIP5 archive. Heavily used.	Deployment	Link vocabularies
Regional climate models	Deployed for the CORDEX archive.	Deployment	Link vocabularies
Statistical downscaling	Development of an agreed set of terms defining methods is in progress.	Conceptual framework	Implementation is supported by IS-ENES2
Global re-analysis	Parameter description vocabularies (in GRIB) are stable and fully implemented; minimal additional work needed to provide additional vocabularies needed.	Mixed:	Create some pilot vocabularies, following on from work on regional re-analysis. Support community effort to link GRIB and CF NetCDF vocabularies.
Regional re-analysis	Developed in collaboration with UERRA. Core concepts well defined; some details to clear up.	Implementation plan	
Satellite observations	Developed in this document, based on well formed data specification of the ESA-CCI programme.	Deployment	Configure compliance checking software and ESGF publication system.
In-situ observations	Work still to be done.		Develop an approach to vocabularies, exploiting well formed EO vocabularies where appropriate.
ISIMIP			
Indicators	Inventory developed in Deliverable 7.1 provides a basis.		

3.2 Categories of categories

This section provides an analysis of the different types of categories used for vocabularies in different data collections.

Different approaches to governance

Governance	Description	Examples
Standards	Clear and effective governance; terms added to meet community requirements.	GCOS ECVs, CF standard names.
Registries	Terms added to suit needs of specialist teams: registered to avoid duplicates and make definitions available. Sometimes with guidance on styling of names.	ESA CCI Product.
Extensible	Allow individual data providers to add terms within a specified pattern.	Ensemble number; file version

Table 1: Different governance approaches to vocabularies.

Different properties

Objective	Description	Examples
View (many worlds)	There are many estimates of the state of the world, and, from simulations, many realisations of possible worlds.	Model, experiment, ensemble; mission, analysis method;
Focus (place within the world)	In each scientific domain there are a variety of ways of identifying focussed areas of information within an overall view of the world; (including both focus on a particular property and on a particular place).	Realm, variable, ECV. Location (and masking), time period.
Style (presentation)	Data can be presented in different ways, varying sampling rate, file format,etc.	Frequency; EO processing level. Spatial mesh. Data format.

Table 2: Vocabularies can be grouped according to different classes of objectives: some aim at distinguishing between different approaches to representing the same thing, others aim at distinguishing between different things.

Different target usage

Vocabularies can be used to label and organise data products, and to help users find data products. Some of the categories in use for the organisation of data products do not have a descriptive definition, they simply act as a short-hand for the list of products assigned to a category. This serves the archivist well, as there is an unambiguous list to define where products belong, but it is of little use for users who are searching for data (e.g. someone looking for snow cover data would need to know that this is in the “meteorology” section of GRIB2 but in the “land” section of CMIP). When asking questions users will be using their own, non-technical language in which terms have broad and overlapping meanings. The vocabularies which are used to organise data, on the other hand, are designed to have clear, if somewhat arbitrary boundaries.

3.3 High level categories.

It is common practice to create broad categories to help the users and managers to deal with the huge variety of physical variables. Different groups have, however, different approaches to this high level categorisation.

GCOS Essential Climate Variables: domains and sub-domains

GCOS specifies 3 domains: atmospheric (split into surface, upper air and composition), terrestrial and oceanic (split into surface and sub-surface) – [not immediately clear how deep the SST ECV goes – possibly get more details from BAMS http://journals.ametsoc.org/doi/pdf/10.1175/2014BAMSStateoftheClimate.1]. (lakes are part of the terrestrial domain).

WMO GRIB: disciplines

GRIB has 7 disciplines: meteorology, oceanography, hydrology, land surface, space weather, and other space. GRIB Hydrology is concerned with surface water: soil moisture is in the land surface section and water in the atmosphere is in meteorology; cloud top height occurs in the space products section …. (it is also in the meteorology section) [what is the difference …]

CMIP Realms

The data specifications for the WGCM Climate Model Intercomparison projects (CMIP) are classified by realms: atmos, ocean, land, landIce, seaIce, aerosol, atmosChem, or ocnBgchem.

These realms are defined by usage rather than by specific rules delineating their scope, and variables may be associated with more than one realm.

Aristotle's elements

The elements of Aristotle (earth, air, water and fire) have stood the test of time and are still used in the Earth sciences. It would be tempting to associate Aristotle's Earth, air and water with land, atmosphere and ocean, but Aristotle's definition is significantly different from those above. Aristotle was interested in material, so that clouds, being mainly water, where considered as part of that category rather than being in the atmosphere.

Aristotle's element of Fire has a broader meaning than our modern interpretation of fire, as can be seen from this quote from a translation of Aristotle's Meteorology: "Some of the vapour that is formed by day does not rise high because the ratio of the fire that is raising it to the water that is being raised is small."

The elements of Aristotle are interesting here as they provide a different perspective, and one which has been both influential and persistent.

Other

There are also the INSPIRE codelists (http://inspire.ec.europa.eu/metadata-codelist/TopicCategory/) used a lot in many geospatial catalogues (e.g. the Sat Apps Cat: http://data.satapps.org/).

The UKEOF Catalogue, as another example, uses "environmental domains" , see e.g. http://onto.nerc.ac.uk/EF/freshwater.html

Discussion

The elements of Aristotle have stood the test of time, and are widely used in the Earth Sciences as a basic categorisation. One limitation is that Aristotle was describing the physical world – the biological world does not enter in.

The three top level categories of the GCOS system (atmospheric, oceanic and terrestrial) have a superficial

resemblance to Aristotle's air, water and earth, but there are major differences: the GCOS atmospheric category includes water vapour and aerosol which belong in Aristotle's spheres of water and earth. The GCOS terrestrial category includes aspects of the biosphere.

The spheres of Aristotle are defined in terms of intrinsic properties, whereas the other classifications reflect, to varying degrees, the priorities of the particular disciplines. The aim here is not, however, to identify a "best" approach, but rather to understand the differences between the approaches in use and ensure that the data infrastructure can mitigate potential confusion arising from varying usages of common words.

The relationships between terms will be expressed a formal relation using the SKOS vocabulary (Appendix 2) and a comment to explain the why that relationship has been declared. The list of relationships will be version controlled to allow managed updates.

e.g.

- ID: 3.01
- Term one: cmip:atmos
- SKOS relation: closeMatch
- Term two: grib2:meteorology
- Comment: The CMIP atmos realm is a process based view of the atmosphere, which closely matches the meteorology, but there are differences: e.g. snow cover is considered as a meteorological product in GRIB, but a "land" product in CMIP.

An initial set of such mappings has been created in a workbook[1]. The NERC Vocabulary Server will provide a means of exploiting these relationships – but does not directly support maintenance and development of the collection of relationships.

This approach allows the task to be dealt with in a structured way. It is proposed that variable level mappings between CMIP and GRIB be dealt with in a similar way, with the help of a web interface to support flexible viewing of the relevant definitions.

Some initial implementations from ECMWF and SMHI will be imported into this framework.

3.4 The CLIPC Book of Terms

The CLIPC book of terms will bring together definitions of a broad range of terms in a common format. This compendium will not seek to replace existing reference sources which are under the control of a range of governance bodies, but will provide a single point of access for key information. The different objectives and scopes of the vocabularies leads to differences in the technical information accompanying each term. Achieving a common format across the whole compendium may lead to some loss of detail.

It will be important to have a clear system for importing updated source vocabularies and flagging relations which may need to be adjusted if the definitions of terms in the source vocabularies is modified.

The greatest challenge is in parameter names, and here CLIPC will support community efforts to define and maintain a system of mappings between GRIB and CF NetCDF. To support users coming to the data with no technical knowledge, links between general terms with domain independent definitions (e.g. "heat wave", "temperature", "downpour") and technical vocabulary terms will be developed.

Decision on level of effort to put in this will be based in part on feedback from user requirements workshop.

Ambiguities in definitions

In many cases the definitions are ambiguous or clear only in the context of the defining domain. For example, the GRIB 2 term "Net long-wave radiation flux [0:0:5]" does not have a specified direction – the flux might be measured positive upwards or downwards, but the API has the clarification that the convention is downwards. For data which complies with this informal convention there is an exact match with the CF standard name "downwelling_longwave_flux_in_air", but if we go by the definition alone it is only a "broaderMatch".

1 http://www.clipc.eu/media/clipc/org/documents/Workbooks/clipc_vocabhlmappings.xls

3.5 Reference point(s) (for citation and curation)

Providing appropriate support for data citation is an essential component of data management. For scientists the key organisational unit for data is a conceptual block which is associated with a single citation. This causes some difficulty in interoperability because the shapes of the conceptual blocks that the scientists expect to see are different in different domains. In the Earth Observation domain we see organisation around teams specialising on a particular set of variable or particular instruments. In the climate modelling domain we see organisation around intercomparison projects and modelling centres. In order to support these different usages the vocabularies have to be agnostic about the reference points.

For curation purposes it is convenient to align archive storage structure with the reference points, which will ideally be associate with Digital Object Identifiers.

Type of activity	Collection of citable actions .. usually associated with a review paper.	Specific citable collection of data.
Climate models	Activity	Model/Experiment group
Re-analysis	Analysis centre	Project
Mobile instruments	Campaign/Measuring Period	Instrument
Space instruments	Platform or Instrument collection	Instrument/Algorithm
Fixed instruments	Instrument	Measuring period

4 Back to basics

This section reviews some terminology which is in common usage across physical sciences and which could be used to organise the vocabularies in a domain independent manner. There are nevertheless some subtleties.

4.1 Basic properties of interest

We might start by defining the basic properties which need to be described. Entities such as air, water, earth. Atmospheric temperature can be taken as the temperature of the air for those parts of the atmosphere which are in thermal equilibrium, but atmospheric radiation is not generally considered as a property of the air. Because of this subtle difference between “air” and “atmosphere” we need to include both in our classification scheme and accept that they are interchangeable in some contexts but not in others.

Common concepts (representing an identifiable state or object): aerosol, cloud (meteorological, dust, chemical), wave, ice-sheet, sea-ice.

Species and or groups of plants and animals (BODC has such a vocabulary for marine species).

The list should be developed from a user perspective.

4.2 Descriptive vocabularies

In addition, there are descriptive vocabularies:

Quantity: amount, concentration, tendency, flux, rate, height, effective radius, mass fraction.

Thermodynamics: temperature, pressure, density, energy, entropy;

Kinetics: velocity, momentum, kinetic energy;

Phase: gas, liquid, solid, solution, suspension;

Location/bin: model lowest layer, EU, particle size range;

Complex measures (need additional qualifiers); reflectivities; condensation rates, phase transitions;

Process categories: “primary source”, “re-emission”;

Macro-properties (shape, texture): roughness, wave heights;

Categorisation: (e.g. nitrate aerosol), cloud types, longwave vs. shortwave (or radiation);

Statistic: e.g. Daily maximum, monthly mean, standard deviation;

5 Conclusions and outlook

The broad scope of CLIPC spans domains with independent and well established vocabularies. Consequently, it is clear that an active system of linking terms in different vocabularies will be needed, The initial focus will be on supporting community efforts to link GRIB and NetCDF vocabularies.

From a user perspective, it is also important to link technical vocabularies (which may, for instance, be constrained to only have one term for each concept) to real world vocabulary in which individual words have broad and imprecise meanings. The SKOS framework for recording relations between terms provides a means of recording and using such a link.

Building and maintaining links between extensive and evolving vocabularies will, however, need a information management work-flow which goes beyond SKOS.

The Food and Agriculture Organisaton (FAO) has an extensive SKOS vocabulary (http://aims.fao.org/aos/agrovoc/) with many environmental terms.

6 Appendix 1: Vocabulary definitions

6.1 Inventory of vocabularies

This section lists definitions used in CMIP5 and additional definitions added or in discussion for subsequent data collections.

		Definition
1.01	CF Standard Names	A name from the CF Standard Name table[2].
	Introduced in CMIP5	
2.01	Activity	The model inter-comparison activity or other data collection activity. For CMIP5 all the archived data will be discoverable under the “CMIP5” activity. For “Transpose AMIP”, the data will be archived under the “TAMIP” activity. In some cases there may be other activities (e.g., CFMIP and PMIP), which have been coordinated with CMIP5, so these activities may be cross-referenced or aliased with CMIP5 for certain portions of the CMIP5 archive.
2.02	Product	Product currently has four options: “output”, “output1”, “output2”, and “unsolicited”. For CMIP5,files will initially be designated as “output”or “unsolicited”. Subsequently, data from the requested variable list will be assigned a version (see below) and placed in either “output1” or “output2. Variables not specifically requested by CMIP5 will remain designated unsolicited”. In some

2Not a great definition, but such circular definitions are unfortunately common in informatics.

		cases a continuous sequence of model data will be split between “output1” and “output2” in order to facilitate archive management. Note that although output of some variables is requested only for limited time-periods, if output of those variables is made available for other time periods, it will also be treated as “output”, not as “unsolicited. It is likely that various data products derived from this output will be produced subsequently which could be identified by a different term (e.g., “derived” or “processed”), but this is not part of the current DRS.
2.03	Institute	Institute identifies the institute responsible for the model results (e.g. UKMO), and it should be as short as possible. For CMIP5 the institute name will be suggested y the research group at the institute, subject to final authorization by PCMDI. his name may differ somewhat from the official CMIP5 institute_id (recorded as a global attribute in CMIP5 output files), which should be used to identify models in journal articles. [The official institute_id might, for example, include characters such as a blank, a period, or a parenthesis, which are not allowed in the DRS “institute” component.]
2.04	Model	Model identifies the model used (e.g. HADCM3, HADCM3-233). Subject to certain constraints imposed by PCMDI, the modeling group will assign this name, which might include a version number (usually truncated to the nearest integer). This name may differ somewhat from the official CMIP5 model_id (recorded as a global attribute in CMIP5 output files), which should be used to identify models in journal articles. [The official model_id might, for example, include characters such as a blank, a period, or a parenthesis, which are not allowed in the DRS “model” component.] The model identifier will normally change if any aspect of the model is modified (e.g., if the resolution is changed). An exception may be made if the modifications to the model are clearly implied by the experiment design. If, for example, a coupled atmosphere-ocean model performs an AMIP simulation (which clearly implies prescribed SSTs and sea ice, rather than a fully interactive ocean), then the name may not necessarily be modified. Another exception is when closely-related “perturbed physics” versions of a model are run, in which case the different model versions can be uniquely identified by assigning each a different “p” value in defining the “ensemble member” (described below).
2.05	Experiment	Experiment identifies either the experiment or both the experiment family and a specific type within that experiment family. In CMIP5, for example, “rcp45” refers to a particular experiment in which a “representative concentration pathway” (RCP) has been specified which leads to an approximate radiative forcing of 4.5 W m-2. As another example, “historicalGHG” is a simulation of the historical” period, but with forcing other than anthropogenic “greenhouse gas” forcing suppressed. In this latter case, “historical” is the experiment family and “GHG” is used to designate the specific type of historical run. These experiment names are not freely chosen, but come from controlled vocabularies defined in the Appendix 1.1 of this document under the column labeled “Short Name of Experiment”. Note that in some cases there will be slight variations of the same experiment (e.g., different simulations performed within the historicalMisc family might be forced with different individual forcings or suites of forcings, as discussed further under “Ensemble member” below)
2.06	Frequency	Frequency indicates the interval between individual time-samples in the atomic dataset. For CMIP5, the following are the only options: “yr”, “mon”, “day”, “6hr”, “3hr”, “subhr” (sampling frequency less than an hour), “monClim” (climatological

		monthly mean) or “fx” (fixed, i.e., time-independent). These are specified for each variable in the “standard_output” spreadsheet found at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Note that for CMIP5, quantities derived from an atomic dataset of a given frequency will be assigned the same frequency, even in the case when a time-average has been performed. (See example under section 2.4 involving time averages.)
2.07	Realm	Modeling realm indicates which high level modeling comp onent is of particular relevance for the dataset. For CMIP5, permitted values are: “atmos”, “ocean”, “land”, “landIce”, “seaIce”, “aerosol” “atmosChem”, ocnBgchem (ocean biogeochemical). These are specified for each variable in the “standard_output” spreadsheet which can be accessed at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Note that sometimes a variable will be equally (or almost equally relevant) to two or more “realms”, in which case the atomic dataset might be assigned to a primary “realm”, but cross-referenced or aliased to the other relevant “realms”.
2.08	Variable	Variable name and the MIP table component of the DRS (defined next) identify the physical quantity and often imply something about the sampling frequency and modeling realm. For CMIP5 the variable name and MIP table for requested output appear in the “standard_output” spreadsheet available at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Monthly mean surface air temperature, for example, has a “variable name” of “tas” and is found in the “Amon” MIP table. Note that hyphens (-) are forbidden in CMIP5 variable names.
2.09	MIP Table	MIP table: See description under the “variable name” component directly above. For CMIP5 each MIP table contains fields sampled only at a single frequency (although in the case of monthly mean data the DRS will place some of the monthly means in the “mon” DRS frequency category and others in the monClim DRS frequency category, as appropriate).
2.10	Ensemble	Ensemble member (r<N>i<M>p<L>): This triad of integers (N, M, L), formatted as shown above (e.g., “r3i1p21”) distinguishes among closely related simulations by a single model. All three are required even if only a single simulation is performed.
2.11	Temporal range	Indicative start and end time (rounded out to a frequency dependent precision).
2.12	Experiment Family	A group of experiments with a shared scientific objective introduced for the user interface, not used in the data files).
	Introduced in CORDEX (dynamical downscaling)	
3.01	Domain	The domain used by the regional model. The definition of the regions includes the technical specification of details of the numerical grid.
3.02	Realisation	Introduced to represent the different configurations etc which might be used with a given model and set of boundary conditions.
3.03	RCM Version Id	Similar to te CMIP5 ensemble element. But defined as: “RCMVersionID (free string; rcm_version_id)identifies reruns with perturbed parameters or smaller RCM release upgrades, i.e. equivalent simulations. Major upgrades and

		improvements should be reflected in the RCMModelName."
3.04	Driving Model	The model used to provide the boundary conditions for a regional model. This term combines the model id and the institute id used to identify the large scale data used.
	Introduced in CORDEX (statistical downscaling) – under discussion – based on March 2013 version.	
4.01	Driving data tracking ids	A list of the tracking ids of the data used to drive the downscaling run.
4.02	Driving Model ID	Very similar to Driving Model element of CORDEX dynamical downscaling, but without the institution id[3].
4.03	Region	Superficially the same as the region element for the CORDEX dynamical downscaling, but the former is defined in terms of a list of predefined regions specified at an early stage of the CORDEX programme. The statistical downscaling approach allows contributors to use region specifications from
4.04	Region Lexicon	Resource locators for lists of valid regions.
4.05	Resolution identifier	A string of the form "hnXXXX" or "hiXXXX" where XXXX is the nominal horizontal resolution of the downscaled data, expressed in kilometers (rounded to the nearest km with leading zeros dropped). "hn" indicates that the data is stored on the model's "native" grid, while "hi" indicates that the data has been interpolated from a model's native grid to a different grid. (Statistically downscaled data would normally be recorded on a so-called "native" grid.) Data on a native grid at a nominal resolution of 5 km, for example, would be identified as "hn5", while regridded data at 11 km resolution would be identified as "hi11". The XXXX should be calculated as follows: XXXX = sqrt(domain area / (number of grid cells)), expressed in km/grid cell and rounded off to the nearest km.
	Introduced in SPECS	
5.01	Experiment family	Experiment_family: Short name of the experiment family. By experiment family is meant the name of the experiment, but without the start date, to avoid hugely complicated names. The official short names of the experiments can be found in the experiment section of the SPECS wiki.
5.02	Start date	Start_date: Start date of the experiment in the form of "Syyyymmdd". For uninitialized runs, a start date will still need to be used. This date corresponds to the beginning of the historical run from the parent experiment (e.g., the pre-industrial simulation). We refer to the date when a historical run starts from the spin up simulation or, more common in our case, the date when the forecast is initialized.
	Introduced in obs4mips	
6.01	Source type	source_type: a character string indicating the intrinsic nature of the data, with currently allowed values of 'satellite_retrieval', 'satellite_merged', 'in-situ', 'ground_retrieval', or 'reanalysis'. We anticipate this might be used in future MIPs with values like 'AGCM', 'OGCM', 'AOGCM', 'EMIC', 'RCM', or 'ESM'. (If

3 There does not appear to be any reference to the institution responsible for the driving data in the March 2013 version of the CORDEX statistical downscaling data specification.

		using CMOR pass this in a call to cmor_set_cur_dataset_attribute.)
6.02	MIP specifications	mip_specs: a space-separated list indicating which model intercomparison project(s)' output specifications have been followed. For example, a dataset that is meant to mimic CMIP5 model output would be assigned the value "CMIP5". (If using CMOR pass this in a call to cmor_set_cur_dataset_attribute.)
6.03	Data structure	data_structure: a character string indicating the internal organization of the data with currently allowed values of "grid", "station", "trajectory", or "swath". The "structure" here generally describes the horizontal structure and in all cases data may also be functions, for example, of a vertical coordinate and/or time. (If using CMOR pass this in a call to cmor_set_cur_dataset_attribute.)
	Introduced in ESA CCI	
7.01	Sensor	Sensor name, e.g. AATSR, use names from CCI common vocabulary. Separated by commas if more than one).
7.02	Platform	Satellite name, e.g. ENVISAT, use names from CCI common vocabulary. Separated by commas if more than one).
7.03	CCI Project	This corresponds to GHRSST in the GHRSST file-naming convention.
7.04	Processing level	e.g. L0, L1 etc
7.05	Indicative date	Date of data, c.f. "Time Range" in CMIP5.
7.06	Indicative time	Time of data, c.f. "Time Range" in CMIP5.
7.07	Data type	A short term describing the main data type in the data set from the list in the CCI Guidelines for data producers. c.f. "variable"
7.08	Data specification version	Including the version number of the GHRSST Data Specification is optional for the CCI filenaming convention. If used it shall be "02.0".
7.09	File version	File version number in the form "n{1,}[.n{1,}]" (that is 1 or more digits followed by optional . and another 1 or more digits).
7.10	Product string	Each ECV team shall define the Product Strings they will use for their data and make this information available in their documentation. The Data Standards Working Group will collate this information to make it easily accessible to data users. The Product String field must not include any hyphens but can include underscores. e.g. "AATSR"
7.11	Additional segregator	This is an optional part of the filename. It must be used if otherwise different data sets would generate the same filename. It can also be used to include in the filename information that doesn't fit elsewhere in the filename convention, but which projects feel is useful for easy identification of different data sets. Each ECV team shall define the Additional Segregators they will use for their data and make this information available in their documentation. The Data Standards Working Group will collate this information to make it easily accessible to data users. More than one element may be included, separated by an underscore, not a

		hyphen. e.g. “DM” for AATSR data (demonstration, c.f. LT: long term, GMPE: GHRSST multi-product ensemble).
7.12	ECV	Not explicitly used, as each project is assigned a single ECV. However, ECV and “ECV project” are distinct from a long term curation perspective, as there may eventually be multiple projects contributing to each ECV.
	ISIMIP	
8.01	Impact sector	e.g. Agriculture, Ecosystems, Health, Water
8.02	Impact model	
8.03	Social forcing	
8.04	CO2 forcing	
8.05	Irrigation forcing	
8.06	Crop	(not used for all cases)
9	ERA-CLIM (discussions in progress – no vocabularies defined yet).	
	GRIB code	GRIB codes specify variables in much the same way as the standard names and CMIP variable names. There is a need to create a mapping from GRIB variables to standard names and MIP names.
10	EURO4M (work to be done within CLIPC).	
	The EURO4M regional re-analysis data will pave the way for UERRA data.	
	RADrivingName	An identifier of the driving ReAnalysis (RA) which can be global or regional one.
	RADrivingEnsembleMember	Identifies the ensemble member and sequence of the driving RA
	Domain	As CORDEX region, with more precise specification of acceptable region lexicons.
	RRAModelName	The model generating the data.
	RRAEnsembleMember	Regional re-analysis ensemble member
	mip_specs	Indicates which data specification has been used (which MIP tables?).
11	Indicators	

11.1	ETCCDI terms	Definitions and short names for 27 Tier 1 indicators.
11.2	D7.1	List of terms generated by CLIPC
11.3	Tier	3 tiers – introduced in CLIPC DoW.

6.2 Notes on vocabulary implementations

In the MIP family of projects, file names are composed of terms separated by an underscore "_", underscores are not allowed in terms; a hyphen is the recommended separator within terms. In the ESA-CCI, terms in the file names are separated by a hyphen "-", and hyphens are not allowed in the terms; an underscore is the recommended separator within terms.

In the MIP family (CMIP5, CORDEX, SPECS, CCMI), the "activity" attribute carries the project name, but this name is not represented in the file name: it can only be deduced by examining the file meta-data or the full path, if available.

The vocabularies listed do not have to be used directly as search facets in the ESGF search interface – it is possible to combine or split terms

6.3 Properties of vocabularies

Some vocabularies have explicit rules about the structure of the terms.

	Character restrictions of values	Provenance of definition
CF Standard Names	Standard names consist of lower-letters, digits and underscores, and begin with a letter. Upper case is not used. [a-z][a-z0-9_]*	
CMIP names	Upper and lower case letter. [a-zA-Z]*	

7 Appendix 2: Requirements for Book of Terms

List of terms, definitions and cross-references.

Editable: require dual search and options for creating links between terms. For example, it should be possible to search for and select a CF NetCDF term, and then search for a GRIB term and after finding both, propose a link between them.

The book of terms will be formulated as a simple relational database, with variables organised in tables. One table will provide definitions of the attributes. For example, NetCDF standard names have 3 attributes: definition, units and alias (optional).

Schema

Section: Vocabulary Definitions.

For each vocabulary specify an identifier (alphanumeric string, no spaces), a title and a description., together

with an optional list of attributes which may or should be specified for each element in the vocabulary. Each attribute must have an identifier (alphanumeric string, no spaces), a title and a description. Note that this schema will not be imposing constraints on the vocabulary elements: any required constraints should be checked before harvesting into this generic format.

Section: vocabulary entries.

A list of terms, each with an identifier (alphanumeric string, no spaces), a description and attributes as specified in the definition section.

Relations and annotations.

Should have a provenance tag (consistent with the PROVO standard) and a description. There should also be a status (proposed, approved, under evaluation etc, with "approved" further specified in terms of authority (e.g. CLIPC approved vs. approval by a community consultation).

8 Appendix 3: Existing Vocabularies

8.1 Essential Climate Variables

The ECVs are an evolving list of core climate variables defined in terms of importance and feasability. GCOS (2010) specifies a list of 50 variables (see below). Bojinkski et al. (2014) discuss the issues around the use and definition of essential climate variables, and the crucial role they play in harmonising planning between agencies. Essential Climate Variables are specified only by a suggestive name, e.g. "Lakes". The specific parameters to be used to represent this variable are not specified.

Domain	GCOS Essential Climate Variables
Atmospheric (over land, sea and ice)	**Surface:** Air temperature, Wind speed and direction, Water vapour, Pressure, Precipitation, Surface radiation budget. **Upper-air:** Temperature, Wind speed and direction, Water vapour, Cloud properties, Earth radiation budget (including solar irradiance). **Composition:** Carbon dioxide, Methane, and other long-lived greenhouse gases, Ozone and Aerosol, supported by their precursors.
Oceanic	**Surface:** Sea-surface temperature, Sea-surface salinity, Sea level, Sea state, Sea ice, Surface current, Ocean colour, Carbon dioxide partial pressure, Ocean acidity, Phytoplankton. **Sub-surface:** Temperature, Salinity, Current, Nutrients, Carbon dioxide partial pressure, Ocean acidity, Oxygen, Tracers.
Terrestrial	River discharge, Water use, Groundwater, Lakes, Snow cover, Glaciers and ice caps, Ice sheets, Permafrost, Albedo, Land cover (including vegetation type), Fraction of absorbed photosynthetically active radiation (FAPAR), Leaf area index (LAI), Above-ground biomass, Soil carbon, Fire disturbance, Soil moisture.

Stephan Bojinski, Michel Verstraete, Thomas C. Peterson, Carolin Richter, Adrian Simmons, and Michael

Zemp, 2014: The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy. *Bull. Amer. Meteor. Soc.*, **95**, 1431–1443.

doi: http://dx.doi.org/10.1175/BAMS-D-13-00047.1

GCOS (2010): IMPLEMENTATION PLAN FOR THE GLOBAL OBSERVING SYSTEM FOR CLIMATE IN SUPPORT OF THE UNFCCC
http://remotesensing.usgs.gov/ecv/document/gcos-138.pdf

8.2 GRIdded Binary (GRIB)

The GRIB standard is designed for data interchange. This has some design consequences. e.g. each record in a GRIB file provides an independent and fully characterised piece of data. Data is encoded as a string of bits, and a large collection of tables provide the interpretation for these bits. There is a transition under-way between GRIB1 and GRIB2. In GRIB1 the range of variables defined in the standard tables is limited and many centres have taken recourse to local tables to encode additional data products. By design such local tables, as the name suggests, should only be used for data held and used internally, but practise does has not followed design. The existence of multiple local tables leads to a loss of interoperability.

The GRIB2 API also provides a short name associated with a parameter triplet or triplet plus a bundle of qualifying parameters. The short names are provided for convenience, but are not part of the standard. In some cases the API adds information, e.g. the GRIB tables do not specify the direction in which radiation fluxes are measured. In the API documentation the comment “by model convention downward fluxes are positive” is added to provide clarification.

GRIB2 is designed to avoid the problem of local tables by having a far greater range of variables in the standard tables and a streamlined process for adding additional variables to these tables. It should be noted, however, that the ECMWF GRIB API software provides a substantial list of variables which are encoded with local tables in GRIB2.

ECMWF is developing an extensive list of mappings from GRIB1 to GRIB2, and mappings from GRIB2 to NetCDF.

The tables of most interest here are “Code table 4.0 - Product definition template number”, “Code table 4.1 - Parameter category by product discipline” and “Code table 4.2 - Parameter number by product discipline and parameter”, which list the discipline, category and parameter names respectively. The triplet of discipline, category and parameter number may give a complete specification of a physical variable, but many variables are represented by a triplet along with additional qualifying parameters such as, for example, a period of time over which measurements have been accumulated.

Many users will interact with the GRIB Application Programming Interface (API) rather than directly with the GRIB format and tables. The GRIB API includes addition short names associated with bundles of parameter settings. These short names are not part of the standard, but provide a convenient labels for users.

The combination (code,short name) is unique for each record – there are 355 such records. There are differences in scope. For instance GRIB2 has specifications for 2 trace gases, CF standard names covers many more. Both systems require explicit registration of new terms.

8.3 Climate and Forecast Conventions (CF)

The CF conventions (http://cfconventions.org/) are primarily concerned with encoding information in NetCDF files, but the convention includes an extensive table of variable names, the “standard name table”, which can be used outside NetCDF files as well known labels for variables.

8.4 Simple Knowledge Organization System (SKOS)

The SKOS terms will be used to provide semantic relations between existing vocabularies. The first 5 listed here are generally used to specify relations between terms within a vocabulary, while the remaining terms are used to express relations between independent vocabularies.

SKOS term	Explanation
narrower	Used to assert a direct (i.e., immediate) hierarchical link between two SKOS concepts
broader	Inverse of “narrower”
narrowerTransitive	Used to record indirect hierarchical links. By convention, nit used to assert (i.e. define) links.
broaderTransitive	Inverse of “narrowerTransitive”
related	Used to assert a relation.
closeMatch	The “Match” options are considered as “mappings” rather than “semantic relations”. By convention these are used to express relationships between concepts in different SKOS concept schemes. “closeMatch” indicates that two concepts that are sufficiently similar that they can be used interchangeably in **some** information retrieval applications.
exactMatch	Concepts can be used interchangeably across a wide range of information retrieval applications.
broadMatch	Hierarchical mapping relation
narrowMatch	Inverse of “broadMatch”
relatedMatch	Used to state that there is some association.

[X broader Y ==> means that Y is broader than X].

A proof-of-concept set of mappings is provided in an Excel workbook here:
http://www.clipc.eu/media/clipc/org/documents/Workbooks/clipc_mappings_poc01.xls

8.5 Provenance terminology: PROVO

PROV-O (http://www.w3.org/TR/prov-o/) provides a standard format for recording provenance.

A small set of provenance terms will add transparency and traceability to CLIPC vocabularies and mappings between vocabulary terms. E.g. for encoding a statement that a relationship between a GRIB API term and a CMIP variable had been defined by SMHI as part of their WP6 activity, we would define the relationship as a “prov:Entity” with “prov:wasGeneratedBy=SMHI”, “prov:generatedAtTime=2012-04-03T13:35:23Z"

and “prov:qualifiedGeneration” assigned to a “prov:Generation” element with “prov:activity=CLIPC (WP6)” and “rdfs:comment=Following discussion among WP partners …..” (see example 7 of http://www.w3.org/TR/prov-o/).

A “prov:activity” can be a generic activity (e.g. “illustration”) or a specific activity occurring at a specified time.

The PROVO XML schema allows elements from other namespaces to be embedded in entity and activity declarations – so an entity can be associated with a SKOS concept.

RDF quads (http://www.w3.org/TR/n-quads/) are an extention of the RDF triple store format which allow provenance information to be attached to mappings and other elements of a triple store.

9 References

AGU index terms: http://publications.agu.org/files/2013/01/AGU_index_terms.doc

NASA GCMD keywords: (http://gcmd.gsfc.nasa.gov/learn/keyword_list.html

WMO GRIB2 code: (https://www.wmo.int/pages/prog/www/WMOCodes/Guides/GRIB/GRIB2_062006.pdf)

WMO GRIB2 code tables for parameters (table 4.2): http://www.nco.ncep.noaa.gov/pmb/docs/grib2/grib2_table4-2.shtml .

GRIB2 organises parameters by discipline/category/variable.

Info on GRIB API http://old.ecmwf.int/publications/manuals/d/gribapi/param/

"GRIB 2 NetCDF" workshop at ECMWF: http://www.ecmwf.int/sites/default/files/GRIB-WS-Proceedings.pdf http://cfconventions.org/Data/cf-standard-names/28/build/cf-standard-name-table.html , http://cfconventions.org/Data/cf-standard-names/28/src/cf-standard-name-table.xml

CMIP5 DRS: http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

CORDEX ADD: http://cordex.dmi.dk/joomla/images/CORDEX/cordex_archive_specifications.pdf

Obs4mip required global attributes: http://obs4mips.llnl.gov:8080/requirements?action=AttachFile&do=view&target=obs4MIPs+Global+Attributes+Requirements+v1.1.pdf

ESA CCI metadata requirements: http://46.137.76.174/sites/default/files/CCI_Guidelines_Iss4.2_May2013_1.pdf

ISIMIP simulation protocol: http://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/research/rd2-cross-cutting-activities/isi-mip/isi-mip-fast-track/simulation-protocol

Glossary

CMIP: Coupled Model Intercomparison Project

GRIB: GRIdded Binary

SKOS: Simple Knowledge Organization System

WMO: World Meteorological Organization