EUROPEAN COMMISSION
Community research

Climate Information Portal

CLIPC DELIVERABLE (D -N°: 5.2)

Metadata and controlled vocabularies specification for data, quality control and uncertainties

File name: {CLIPC_52.docx}

Dissemination level: PU (public)

Author(s):
Ruth Petrie (STFC)
Victoria Bennett (STFC)
Martin Juckes (STFC)

Reviewer(s):
Rob Thomas (BODC)
Jan-Willem Noteboom (KNMI)

Reporting period: Dec 2013-Dec 2015

Release date for review: 28 Jan 2016

Final date of issue: 23 Feb 2016

Revision table			
Version	**Date**	**Name**	**Comments**
1.0	17 Dec 2015	First version	Circulated for discussion and internal review
1.1	22 Feb 2016	Final updated version	Updated after reviewer comments and changes accepted.

Abstract

The CLIPC project is building a data services layer to a range of different datasets, located at various data centre and institutes, produced by different science communities. To allow these data to be described, discovered, accessed and used, in a consistent manner, it is necessary to develop cross-data standards for metadata and controlled vocabularies.
This document describes the approaches used, and summarises the work carried out to date, on metadata and controlled vocabulary specification for data, quality control and uncertainties.

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

Table of Contents

Executive Summary

The aim of this deliverable is to describe in detail the use of metadata and controlled vocabularies for data, quality control and uncertainty information within CLIPC.
Since it is the aim of the portal to provide harmonised data search across a variety of different types of climate data records (e.g. climate model simulations, reanalyses and observational data) it is important that the discovery metadata is complete and fully descriptive.

Controlled vocabularies play an important role in ensuring consistency across the different communities involved with CLIPC. Where different vocabularies are required or already exist, mapping relationships between the terms will enrich the data and facilitate the harmonisation of data access.

The wide range of formal and informal conventions in the different communities providing climate data presents users with a bewildering kaleidoscope of formats.

Whilst climate model datasets produced in the Coupled Model Intercomparison Project, Phase 5 (CMIP5) conform to agreed data reference syntax (DRS) and controlled vocabularies, additional datasets considered and included in the CLIPC platform have required these DRS and vocabularies to be developed, and mapped to existing systems. Much progress has been made with the DRS, controlled vocabularies and mappings for a range CLIPC datasets, and in particular the European Space Agency (ESA) Climate Change Initiative (CCI) data.

The metadata for uncertainties across datasets is not consistent. A case study of CCI data products is included here, and a potential way forward is proposed. Further progress on improving consistency and creating standard vocabularies for uncertainty will be made at an Uncertainties workshop in Hamburg in February 2016. It would be beneficial to include additional controlled vocabularies on the quality and commentary metadata, this work will be pursued early in 2016.

Further work is also required on the metadata, DRS and controlled vocabularies for climate impact indicators, for which a workshop is planned in Toulouse in February 2016 to help resolve outstanding issues.

1. Introduction

The objective of this document is to describe in detail the metadata standards and controlled vocabularies rationale and development for CLIPC climate data records. This document builds on the work in previous deliverable and milestone documents (D5.1[1], MS18[2] and MS19[3]).

The CLIPC portal will bring together data from a variety of sources such as in-situ observations, satellite observations, global and regional climate models and global and regional reanalyses. Each of the different sources of data may have their own data types, formats and metadata specifications. For example, satellite data requires metadata relating to the satellite and instrument used whereas in-situ data may require metadata on the location and elevation of the observation.

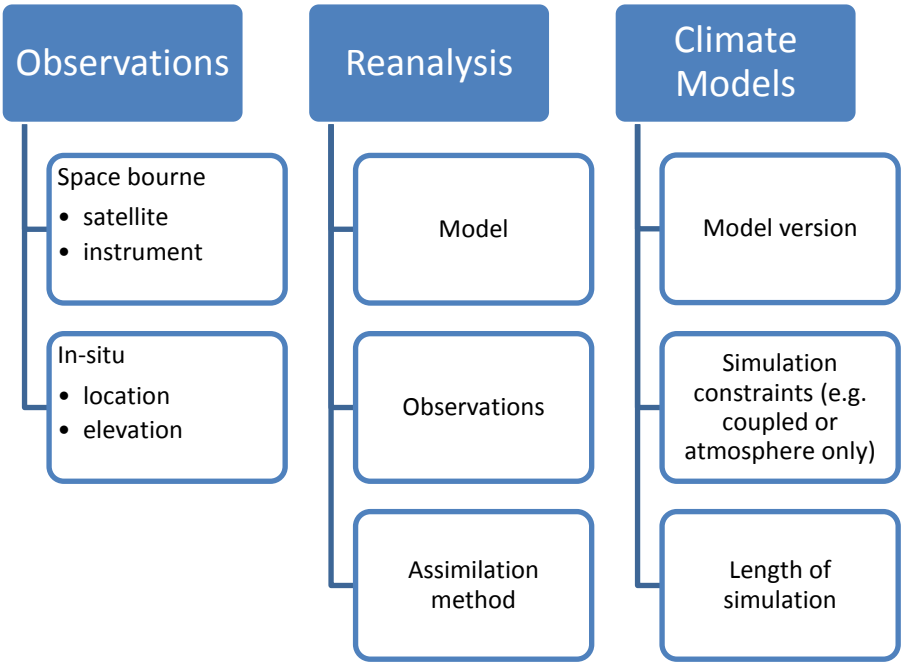

Figure 1: Schematic illustration of a subset of observational metadata

1.1 Metadata, controlled vocabularies and simple knowledge organization systems

In order for data to be discoverable and used in processing and visualisation through the CLIPC portal, the metadata must conform to predefined standards. It is not required that all data conform to one metadata standard, rather a small number of relevant standards have been selected and is expected that all CLIPC datasets should have metadata that conforms to a CLIPC metadata standard. Where new datasets are to be incorporated into the CLIPC portal

1 http://www.clipc.eu/content/content.asp?management=true&menu=0001_000044

2 http://www.clipc.eu/media/clipc/org/documents/milestones/ms18%20extended%20specifications%20for%20climate%20data%2020150429.pdf

3 http://www.clipc.eu/media/clipc/org/documents/milestones/ms19_drsvocabs_april2015_final.pdf

the Metadata Producers (usually the data producer) and Metadata Curators (usually the data curator) will work together to ensure that the appropriate metadata data standards are met. This enables the CLIPC portal catalogue operator to make all data irrespective of source discoverable through the portal. These roles are discussed in more detail in section 4.2. Following Lawrence et al. (2009) metadata taxonomy consists of the following key classes of metadata:

A: Archive metadata. Describes the syntax and semantics (e.g. parameter descriptions) of the data objects themselves.

B: Browse metadata. Supports understanding the context of data and choosing between similar datasets.

C: Character/Commentary metadata. This includes citations of the data itself and post-fact assertions as to the quality of the data.

D: Discovery metadata. This is a subset of the browse and archive metadata, which is selected to aid finding data for evaluation or visualization and/or other uses.

E: Extra metadata. This is the core discipline or instrument-specific metadata, which may be strongly typed (i.e. conforms to schema) or consist of arbitrary documents.

Different classes of metadata are used in different parts of the CLIPC system. Catalogue metadata (also termed discovery metadata) are used to aid discovery of datasets. This catalogue metadata is derived from metadata embedded in the dataset (files), supplemented in some cases by descriptive information supplied separately by the data provider. Thefile-level metadata also provide information about file contents allowing users to select the subsets they require.

A controlled vocabulary is a standardized set of terms that define terminology used to describe and organise data. The Simple Knowledge Organisation System (SKOS) allows relationships between terms in vocabularies to be defined. Using controlled vocabularies and defining hierarchical and associative relationships between them will facilitate data discovery on the CLIPC portal.

Climate data records are enriched by providing quality control and uncertainty information. Metadata and controlled vocabularies when used for this information can further facilitate data discovery, enable consistent data interpretation, and allows users to make informed decisions on the use of the climate data record.

The CLIPC system will include calculation of impact indicators, using the datasets accessible through the platform. In selecting input datasets for these calculations it is important that data are provided in a consistent manner with the appropriate use of metadata and controlled vocabularies. The Tier1, Tier2, and Tier3 indicators that are calculated, and made available through the Platform will, in turn, require descriptions using standardised vocabularies and metadata.

1.2 Data Predicates

Predicates, key, attributes and tags are conceptual modelling terms that are all variations on the entity-relationship theme. The general principle is that a web resource (i.e. a page or file on a web site) can be assigned a set of key/value pairs which facilitate structured search

queries. In the Resource Description Framework (RDF)[4] predicate/object pairs are the only way in which information is attached to a subject.

The CLIPC predicates form a crucial part of the CLIPC vocabularies. The term is used here in the sense in which it is defined in RDF, namely, as a term which expresses a relation between two other terms. In RDF information is expressed in terms of triples of the form "subject, predicate, object", which can be thought of, at least in the context of this document, as equivalent to "record, key, value". Thus, any subject or record is defined by a collection of "predicate, object" (or equivalently, "key, value") pairs (e.g. "model=HadCM3", "standard_name=air_temperature").

1.3 Data Reference Syntax (DRS)

The phrase "Data Reference Syntax" was introduced by Taylor et al. (2012)[5] to describe a collection of conventions designed to ensure that the data in the CMIP5 archive could be clearly referenced, with clear naming rules for every file. Examples of DRS vocabularies include names of climate models and numerical experiments. The CMIP5 DRS vocabularies are used in the file metadata, in the publication workflow, in the documentation, the archive catalogue and the user interfaces. Taylor et al. (2012) define the vocabularies of the CMIP5 DRS and give guidance on their usage.

Within CLIPC we will exploit ideas taken from the RDF to develop a framework which allows vocabularies and their usage to be defined systematically without being dependent on the specifics of applications which might exploit the vocabularies. In these terms, the predicates of the CMIP5 DRS are: activity, product, institute, model, experiment, data sampling frequency, modeling realm, variable name, MIP table, ensemble member, version number and time range. The DRS can be split into three components:

(a) DRS predicates

The predicates are the concepts which are used to define the archive resources. For example, "Sensors" is a CCI vocabulary listing Earth Observation instruments which take measurements from space.

(b) DRS vocabularies

The vocabularies are lists of terms which are valid values of the predicates. For example, "SCIAMACHY" is a term in the "Sensors" vocabulary, referring to an instrument which was flown on the ESA ENVISAT satellite[6]

(c) DRS usage

The usage information specifies how different predicates are used in file metadata, archive catalogues etc. For example, in the ESA CCI data files, the global attribute "sensor" is set to a term from the "Sensors" vocabulary.

In CLIPC, it is planned to incorporate the DRS predicates and vocabulary information into the catalogue metadata using the Simple Knowledge Organisation System (SKOS, see section

4 http://www.w3.org/TR/rdf11-concepts/

5 cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

6 https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat

3.2) Unique Resource Identifiers (URIs). This opens up the potential for catalogue searches using the same "facets" (or predicates) as are used in ESGF's own search functionality. Note that not necessarily all DRS predicates will be included in the discovery metadata.

1.4 Document Outline

In section 2 we describe the CLIPC metadata standards development, in section 3 we discuss the use and development of controlled vocabularies and SKOS: the simple knowledge organization system within CLIPC. In section 4 we discuss catalogue (discovery) metadata. The use of metadata and controlled vocabularies for quality control, uncertainty data and impact indicators are discussed in sections 5, 6 and 7 respectively. Finally, in section 8 we provide some conclusions and plans for future development.

2. CLIPC Metadata standards development

It is now common that data is held within files that have internal structures which allow for significant amounts of descriptive metadata to be embedded within the files. This could be a combination of Archive, Browse and Discovery metadata (Lawrence et al., 2009). Different communities have developed their own file level metadata specifications, typically conforming to an agreed standard (e.g. CF compliant NetCDF), and often with additional terms added to meet user or data provider needs (e.g. ESA CCI data standards include additional metadata fields indicating the satellite sensor and platform).

2.1 File level metadata specifications

The file level[7] metadata of existing mature datasets have been adopted in CLIPC, whilst where datasets are still being created, metadata are being developed to align with the most suitable existing metadata specification.

The table below lists the specifications in use from the Dataset Inventory (D5.1)

Dataset	Metadata specifications for CLIPC publication
EURO4M 3D-VAR reanalysis	Under development (WP5 activity)
EURO4M MESAN reanalysis (downscaled)	Under development (WP5 activity)
ECA&D e-OBS gridded dataset	Under development (WP5 activity)
ESA-CCI	ESA CCI Data Standards
HadOBS	Under development (WP5 activity); likely to be based on ESA CCI Data Standards
EUMETSAT CMSAF - global	Deferred (likely link to EUMETSAT system, not publish through CLIPC)
GlobSnow	Based on ESA CCI Data Standards
CORDEX	Based on CMIP5 specifications
SPECS	Based on CMIP5 specifications

[7] File level here describes the individual files with a given dataset, e.g. the CCI-SST dataset that would contain many files. This is sometimes referred to a data granule or atom. Depending on how the data are organized a dataset can contain a number of granules, which could be a file or a table, view, blob, query, etc.

In general, model data follow specifications based on those developed in CMIP and observational data incorporated in CLIPC will follow specifications similar to those developed for the ESA CCI. Additionally, a number of new datasets will be produced within the CLIPC project; e.g. impact indicators and bias corrected model data. These should, as far as possible, follow one of the metadata specifications mentioned in the table above.

2.2 Data Reference Syntax

Use of a Data Reference Syntax (DRS) within CLIP-C requires that

1. The DRS provides a unique identifier for each dataset.
2. The DRS should provide a clear and structured set of conventions to facilitate the naming of data entities within the data archive.
3. The DRS should make use of controlled vocabularies to facilitate discovery and ensure consistency.
4. The DRS should be consistent with the Earth System Grid Federation (ESGF) framework.

Since CLIPC seeks to harmonise data from a number of different communities, multiple DRS are required. The DRS that are needed for the different types of data within CLIPC are detailed below.

2.3 Satellite Observations (e.g. CCI)

Many satellite datasets are distributed through the Obs4MIPs project[8], which has a carefully designed DRS and metadata specifications which maps closely to CMIP5 specifications. The CCI programme, however, found these specifications to be too restrictive and developed a more flexible approach which is adopted in CLIPC. In particular, the Obs4MIPs protocol did not support the inclusion of uncertainty information in the data files, had only limited support for Earth Observation specific products, metadata fields and parameters, and was incompatible with existing conventions such as the GHRSST convention[9].

The DRS for satellite observations should, where possible, follow the DRS which is under development for the ESA CCI programme. The proposed DRS for the CCI is outlined below, it is a summary of the more detailed version in MS19[10].

Reference vocabulary		Search vocabulary	
Programme	The project producing the data: corresponds to ECV	**Platform**	Satellite(s) from which data have been included in the ECV
CCI project	EO processing level, using standardised CCI definitions.	**Sensor**	Sensor(s) from which data have been included in the ECV
Processing level	Physical variable, e.g. Aerosol optical depth	**Algorithm**	May be indicated in various places.
Data Type	A string sometimes	**Standard**	Specified in file.

10 http://www.clipc.eu/media/clipc/org/documents/milestones/ms19_drsvocabs_april2015_final.pdf

	identifying the product in terms of the input dataset, sometimes referring to an algorithm	**name**	
Product	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.	**Frequency**	Frequency is a key component of the data discovery options for climate model data.
Additional Seg.		**Spatial resolution**	Following the approach set an indicative resolution in km.
Indicative date		**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
Indicative time		**Equivalent CMIP name**	Where possible, a corresponding CMIP variable name.
Data specification version	As used in GHRSST.		
File version	Need to clarify if reprocessing with a new data specification version should also result in an incremented file version number.		

2.3.1 In-situ observational data

The main source of in-situ observational data for use in CLIPC will come from the UK Met Office Hadley Centre "HadOBS" programme[11]. A DRS for this data has been developed by the HadOBS team. The recommended facets are:

DRS for dataset		DRS for filename	
project	CLIPC	**variable**	AMIP style variable names
product	product type: in-situ	**collection**	Dataset
inst	Institute(s)	**framework**	Dataset framework e.g. HadOBS
framework	Dataset framework e.g. HadOBS	**realization**	Required for ensemble of data
collection	Dataset	**project_version**	Native version control
Freq	mon or hr	**start_date**	YYYYMMDD

11 http://www.metoffice.gov.uk/hadobs/

table	Table is a way of being very specific about what is contained in the file – gridbox and temporal boundaries, spacing, averaging – essentially all the stuff that is contained in the file attributes.	**end_date**	YYYYMMDD
realization	Where there is only one realization it will only be r1. For HadCRUT4 it will be r{1-100} and r0 for a median.		
product_version	Native version control		
version	The ESGF version system – usually YYYYMMDD.		

2.3.2 DRS for model data

The DRS for model output should make use of the DRS developed for CMIP5, and to allow inclusion within ESGF the structure should be as follows:

DRS for dataset		DRS for filename	
activity	Modelling activity, e.g. CMIP5 *CORDEX*	**variable name**	
product		**table**	
institute	Institute	**model**	As in dataset
model	Model and its version *Regional driving model*	**experiment**	As in dataset
experiment	Group of experiments e.g. in CMIP - rcp45 *Driving experiment*	**ensemble_member**	Ensemble member reference
frequency	Temporal frequency of output, yr, mon, day, etc	temporal_subset	e.g. YYYYMM-YYYYMM-clim
modelling realm	CMIP5 controlled vocabulary of modelling realms e.g. atmos	geographical info	e.g. g-{region/bounding box}
regional domain	*CORDEX regional domain*	*startTime – endTime*	
table	MIP lookup table		
ensemble member	Ensemble member reference *Driving ensemble member*		
version number	Version of publication-level dataset YYYYMMDD		

In the Coordinated Regional Climate Downscaling Experiment (CORDEX) these facets are slightly modified, where this is the case these are in black italics in the above table[12].

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

2.3.3 DRS for bias adjusted data

A DRS framework for bias adjusted data was developed by SMHI for the CORDEX regional climate model downscaling project (See [13] for full details), a short summary is provided here. Three bias-correction DRS sub-elements are introduced:

BCname (CV[14]): an identifier for the applied bias-correction method that includes a combination of acronyms for the institute and the bias-correction method, separated by dashes "-".

OBSname (CV) is an acronym for the observation/reanalysis datasets used as a reference for bias adjustment. Presently, there is no unique Controlled vocabulary for regional observational datasets, and acronyms for observations have to be defined in consultation with institutions responsible for the observational products.

REFperiod - reference or calibration period in YYYY-YYYY format.

These 3 sub-elements are attached using dashes (-) to the CORDEX DRS element *RCMVersionID* (the 'Downscaling Realisation' search facet in the CORDEX-ESGF segment) creating a new element:

BiasAdjustment : <RCMVersionID>-<BCname-OBSname>-<REFperiod>

The DRS filename should have Adjust appended after the variable name. Within the NetCDF file the long variable names should be prefixed with bias-corrected. Additionally the NetCDF file should have the following global attributes included: bc_contact, bc_method, bc_method_id, bc_observation, bc_observation_id, bc_period, bc_info.

2.3.4 DRS for Impact Indicators

The CLIPC portal will allow for the distribution and calculation of Tier 1, Tier 2 and Tier 3 impact indicators. The climate community has yet to establish a consistent approach to metadata for impact indicators, see MS26[15]. A working group from CORDEX, ETCCDI, ET-SCI and CLIPC has been formed and work towards standardization of metadata for impact indicators is ongoing. Impact indicators will also require appropriate data reference syntax (DRS). It is unlikely that an existing DRS will be able to fully describe the facets of the impact indicators, some additional facets will be needed. These will be made available as soon as possible.

A database of impact indicators has been generated, which includes information about relationships between indicators. This work should be completed soon and will then be available on the early-release version of the CLIPC portal.

13 http://www.cordex.org/images/pdf/guidelines/CORDEX_ESD_Experiment1.pdf

14 From a controlled vocabulary

15 http://www.clipc.eu/media/clipc/org/documents/milestones/clicp_milestone_m26_20151109.pdf

2.3.5 DRS for reanalysis data

A DRS for the Uncertainty Estimates in Regional Reanalysis (UERRA) was developed, it shares many facets of the Coupled Model Regional Downscaling Experiment (CORDEX) DRS[16].

DRS for dataset		DRS for filename	
activity	Project id	**VariableName**	
product	Output	**Domain**	Regional domain identifier
Domain	Regional domain identifier	**RADrivingName**	As in dataset
institute	Institute	**Experiment**	As in dataset
RADrivingName	Driving Model and its version	**RAdriving_ensemble_member**	As in dataset
experiment	Experiment identifier	**RAModelName**	Ensemble member reference
RADrivingEnsmbleMember	Driving ensemble member identifier	**RAEnsembleMember**	
RAModelName	Name of regional analysis model	**Frequency**	yr, mon, day, etc
RAEnsembleMember	Ensmble member identifier	**StartTime-EndTime**	e.g. YYYYMM-YYYYMM
Frequency	Temporal frequency of output, yr, mon, day, etc		
variableName			

A DRS for Global reanalysis would be similar but would refer to global models, domains and full ensemble members and experiments. A DRS for ana4MIPs[17] (a reanalysis intercomparison project) specified a DRS for the data set, although only a subset of reanalysis data, it is given below:

cmor_version	*the version of CMOR that wrote the data if CMOR used)*	**product**	"observations" or "reanalysis", which indicates what type of data you are writing.
contact	name and contact information	**project_id**	"ana4MIPs"
Conventions	CF-version	**realm**	Character string that indicates the portion of the earth system for which the variable is particularly relevant. Using the CMIP5 modelling realms.
creation_date	a character string epresentation of the date when the file was created in the format: "YYYY-	**references**	a list of published or web-based references that describe

16 http://www.ecmwf.int/sites/default/files/elibrary/2014/13713-data-reference-syntax-governing-standards-within-climate-research-data-archived-esgf.pdf

17 https://www.earthsystemcog.org/projects/ana4mips/ProjectDescription

	MM-DD-THH:MM:SSZ		the data or the methods
experiment_id	<source_id>-reanalysis. A short string identifying the model used by the reanalysis	**source**	Character string fully identifying the observational product and version.
frequency	Temporal frequency of output, yr, mon, day, etc	**source_id**	character string containing an acronym that most users would associate with the data product
institute_id	A short acronym describing the 'institution' facet, e.g. might describe funding agency	**table_id**	Character string identifying the CMOR table where this variable appears of the form "Table <table name>"
institution	the institution that generated the data	**tracking_id**	Character string that is almost certainly unique to this file and must be generated using the OSSP[18] utility which supports a number of different DCE 1.1 variant UUID options[19]
mip_specs	a space-separated list indicating which model intercomparison project(s)' output specifications have been followed.	**comment**	Character string containing additional information about the data or methods used to produce it.
model_id	a string containing an acronym that identifies the model use to generate the reanalysis output.	**history**	Character string containing an audit trail for modifications to the original data.
modelling_realm	CMIP5 modelling realm	**title**	A description of the data found in the file.

Terms in italics denote optional facets, green are facets not required by CMIP5, red are facets required by CMIP5 but not reanalysis

3. Controlled vocabularies and SKOS

A controlled vocabulary is a standardized set of terms that define terminology used to describe and organise data. A set of controlled vocabularies can be related using the Simple Knowledge Organization System (SKOS)[20] see section 3.2 (these concepts were described in the CLIPC architecture document[21] and MS9[22]). A short description of how to create a controlled vocabulary is outlined below.

[18] OSSP is a Universal Unique Identifier (UUID) http://www.ossp.org/pkg/lib/uuid/

[19] There are many equivalent libraries to OSSP available, crucially they must comply with the UUID specifications: http://www.itu.int/rec/T-REC-X.667-201210-I/en

20 SKOS - http://www.w3.org/2004/02/skos/

21 http://www.clipc.eu/media/clipc/org/documents/other/clipc_at_v1_1_feb2015.pdf

22 http://www.clipc.eu/media/clipc/org/documents/milestones/ms9%20clipc%20vocabulary%20design%20v2.pdf

Creating a controlled vocabulary

The methodology that has been adopted for producing machine readable vocabularies is as follows. First a series of categories were identified that would benefit from the use of controlled vocabularies. The domain expert then collated a list of terms (concepts) for a vocabulary in a spreadsheet. Each vocabulary was saved in a separate spreadsheet. In order to track changes a version control system (in our case GitHub) was used to store the spreadsheets. The use of spreadsheets was chosen as they are suitable for tabulating data and they are a tool that the domain experts are familiar with. In order to aid the parsing of the data the layout across the spreadsheets was standardised.

A controlled vocabulary spreadsheet consists of a minimum of four columns

- Key: A unique resource identifier (URI)
- Label: A human-readable title of the term, this is also unique
- Alternative label: An alternative label that may be for example, an abbreviation
- Definition: A full description and definition of the term.

Example: Subset of the controlled vocabulary defining the ESA CCI:

URI	Pref Label	Alt Label	Definition
aerosol	aerosol	aerosol properties	Aerosol properties climate data record produced from satellite data as part of the European Space Agency (ESA) Climate Change Initiative (CCI)
cloud	cloud	cloud properties	Cloud properties climate data record produced from satellite data as part of the European Space Agency (ESA) Climate Change Initiative (CCI)
fire	fire		Fire disturbance (burned area) climate data record produced from satellite data as part of the European Space Agency (ESA) Climate Change Initiative (CCI)

Once a spreadsheet is checked into GitHub by the domain expert, the vocab machine admin can take it and convert it into a csv format. The data then undergoes a number of automated transformations. First the csv are parsed and converted into SKOS and then exported to files in Turtle (Terse RDF Triple Language[23]) and xml/rdf formats. The Turtle file is used in the generation of a web page. The Turtle is uploaded to a SPARQL server running on the vocab machine and the Turtle, xml/rdf and html files are transferred to the web server on the vocab machine. This process will be modified in the near future if the decision is made to host the vocabularies on the NERC vocabulary server. Controlled vocabularies have been developed as described above for CLIPC, and progress to date is summarized here.

23 www.w3.org/TR/turtle/

3.1 Controlled Vocabularies for CLIPC

The objective of the controlled vocabularies for CLIPC is to provide a set of predefined terms which can be used in a consistent way to classify datasets. The use of the vocabularies will help with the design and implementation of data portals, by facilitating easier data discovery and hence enhancing the end users experience. Some of the terms may be generic, others specific to a particular activity, group of activities or data.

Scientific domains such as chemistry and biology have well established globally recognised vocabularies. However, there is still a considerable degree of variation in vocabularies used in the Earth System sciences (MS19[24] provides a detailed review of the current extended controlled vocabularies used in Earth system sciences relevant to CLIPC). The vocabularies differ between different communities of the Earth system sciences, e.g. earth observation, climate modelling as each has their own data requirements. Even within each community different vocabularies exist, e.g. global climate modelling and regional climate modelling.

The CLIPC portal will provide access to data from in-situ observations, satellite instruments, climate models and reanalyses. Each different type of dataset will have a different controlled vocabulary associated with it, although they are often already harmonised within their domain, the syntax and semantics can be different. Using standardized vocabularies is an important step in harmonised data discovery and access. Within CLIPC the controlled vocabularies will provide several services:

- Provide definitions for CLIPC project terminology.
- Provide definitions of the search terms in the data discovery service.
- Provide definitions of uncertainty data terminology.
- Provide definitions and documentation of the calculation and processing services implemented in the portal to generate the Tier 1, Tier 2 and Tier 3 impact indicator data products.
- Provide standardarized metadata terminology and commentary terms.

Many controlled vocabularies which describe CLIPC data are already developed and available for use within CLIPC. For example CLIPC will make use of the definitions and hierarchy in the SeaDataNet and the NERC Vocabulary Service (NVS). The NVS can assist in mapping the discovery terms to one single system, to optimise search and discovery, therefore the extension of the NERC vocabulary services is a key development within the CLIPC project. To reconcile the different vocabularies for the different climate data records the Simple Knowledge Organisation System (SKOS[25], see sub-section 3.2) is used to provide a mapping framework that will link the different vocabularies.

24 http://www.clipc.eu/media/clipc/org/documents/milestones/ms19_drsvocabs_april2015_final.pdf
25 SKOS - http://www.w3.org/2004/02/skos/

MS19[26] document presented an inventory of vocabularies and which are summarized here:

Data type	Vocabularies
High level categories	GCOS Essential Climate Variables (domains and sub-domains), WMO GRIB disciplines, CMIP Realms, INSPIRE codelists, UKEOF Catalogue inventory
Variable names	CF Standard names, GRIB codes, EUMETSAT CMSAF
Global climate models	CMIP standards, SPECS, Obs4MIPS ISMIP
Regional climate models	CORDEX definitions
Globabl reanalysis products	
Regional reanalysis	EURO4M (work to be done within CLIPC)
Satellite observations	ESA Climate Change Initiative (CCI) – work done within CLIPC
In situ observations	HadOBS – work ongoing in CLIPC
Impact indicators	Expert team on climate change detection and indices (ETCCDI)

Some additional controlled vocabularies that are not already in existence have been developed for use in CLIPC, a list of the new controlled vocabularies in the following section. The migration the new controlled vocabularies from STFC to the NERC Vocabulary Server[27] (NVS) is under discussion.

3.1.1 New Controlled Vocabularies and SKOS mappings developed for CLIPC

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 the Citation Typing Ontology[28] (CiTO), however it is likely that PROV-O[29] will be used when the new vocabularies are incorporated into the NERC Vocabulary Server (NVS).The new vocabularies include 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. Similar conceptual themes exist for the GRIB disciplines, the CMIP realms the Climate4Impact themes and the ESA Climate Change Initiative (CCI) – ECVs. Using SKOS the internal hierarchical and associative mappings are defined. More detailed information was required for the ESA-CCI, and new controlled vocabularies were defined, and are detailed below.

26 http://www.clipc.eu/media/clipc/org/documents/milestones/ms19_drsvocabs_april2015_final.pdf
27 http://vocab.nerc.ac.uk/
[28] http://www.essepuntato.it/lode/http://purl.org/spar/cito
[29] https://www.w3.org/TR/prov-o/

Controlled Vocabulary	Additional Details
GRIB disciplines	Disciplines for GRIB encoding
CMIP realms	Coupled Model Intercomparison Project modelling realms
GCOS – ECV domains	High level domains in which the GCOS ECVs sit
Climate4Impact themes	To be used as the themes on the portal front page
GCOS – ECVs	Individual GCOS Essential Climate Variables
CCI ECVs	Individual ESA Climate Change Initiative - Essential Climate Variable projects
IPCC Glossary of terms	Terms taken from the IPCC Data Distribution Centre
CCI – Organisations	Organisations involved in the CCI with their registered institute number where appropriate
CCI – Platforms	List of platforms used in the CCI
CCI – Sensors	List of sensors used in the CCI
CCI – Processing Level	List of all processing level acronyms
CCI – frequency	Temporal frequency of CCI data
CCI – CF parameters	The CF standard names used in the CCI

SKOS Mappings	
CMIP realms	Internal hierarchy
GCOS ECV domains	Internal hierarchy
CMIP realms ↔ GCOS ECV Domains	Associative relationships
CIMP realms ↔ GRIB disciplines	Associative relationships
CCI ECVs ↔ GCOS ECV domains	Associative relationships
CCI CF parameters ↔ GRIB codes	Associative relationships
CCI Platforms ↔ CCI Sensors	Associative relationships

All the work on the construction of the controlled vocabularies is available on GitHub (https://github.com/cedadev/cci-vocabularies/tree/master/data). This work is on-going, additional controlled vocabularies for other terms will be added in the coming months. The long term (i.e. post-CLIPC) governance of these new controlled vocabularies is under consideration. It is anticipated that while STFC remains a partner in related NERC projects that STFC will continue to manage these vocabularies, adding new terms where appropriate. It is hoped that future projects will continue to make use of these vocabularies, however if they are not used by the community in the long term or are superseded then it is likely that the vocabularies would be deprecated. There is potential to include new vocabularies for additional glossary terms, for terms relating to the impact indicators or the inclusion of a controlled vocabulary for the CORE-CLIMAX maturity matrix.

3.1.2 A glossary of terms

A glossary of terms taken from the IPCC Data Distribution Centre[30] (DDC) has been incorporated as a new controlled vocabulary. The IPCC-DDC glossary is based on the IPCC-Fifth Assessment Report glossary of terms and a list of appropriate acronyms. It is intended

30 http://www.ipcc-data.org/guidelines/pages/glossary/

that this glossary also be included as a controlled vocabulary in the CLIPC portal. As with the other vocabularies a web page will be available on the vocabulary server. Also where they are available links to the original sources of the definitions are included in the form of cito: citesAsSourceDocument[31].

The Glossary will be used as definition-of-terms list behind the portal (together with the list from Euporias[32]/C4I[33]). It is envisaged that a portal would hold a cache of the terms in the glossary and periodically (daily?) update this cache. This will allow for pop up definitions; whenever a term is used on the website that is in these lists the user will see an underlining of the term, hovering the mouse over the term will then show the definition. In addition the glossary will increase searchable terms.

The controlled vocabularies will be used in several ways:

- To achieve a harmonised search over several data sources. For example, a search on the CLIPC portal for "Sea surface temperature" would be translated to the equivalent CF term; this would then be used to search for products in the CLIPC and MyOcean catalogues. The mapping lookup uses the related P02 term (from the NERC Vocabulary Server) to search SeaDataNet[34] and EMODNet[35] data products.
- To support the metadata descriptions of datasets enabling the use of a controlled term for parameters, instruments, model etc.
- To provide definitions to users of the vocabulary terms in metadata descriptions or search interfaces.

Detailed implementation plans were provided in the CLIPC D3.1[36] deliverable document.

3.2 Simple Knowledge Organization System (SKOS)

The CLIPC Architecture document[37] and MS9[38] describe SKOS in detail; here a brief overview is given. The fundamental element of a SKOS vocabulary is the "***concept***". Concepts can be units of thought, ideas, meanings, or (categories of) objects and events. In the context of the CLIPC data a concept could be e.g., the GCOS Essential Climate Variables (ECV), the CMIP Realms or the CF Standard names.

31 http://www.essepuntato.it/lode/http://purl.org/spar/cito
32 http://www.euporias.eu/
33 http://climate4impact.eu/impactportal/general/index.jsp
34 http://www.seadatanet.org/
35 http://www.emodnet.eu/
36 http://www.clipc.eu/content/content.asp?menu=0001_000042
37 http://www.clipc.eu/media/clipc/org/documents/other/clipc_at_v1_1_feb2015.pdf
38 http://www.clipc.eu/media/clipc/org/documents/milestones/ms19_drsvocabs_april2015_final.pdfhttp://www.clipc.eu/media/clipc/org/documents/milestones/ms9%20clipc%20vocabulary%20design%20v2.pdf

SKOS provides relational matches between two predefined vocabularies using the Resource Description Framework[39] (RDF). The SKOS relationships between different vocabularies are broadly 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 the controlled vocabularies in the SKOS framework allows us to define mappings of relationships between the components of the CLIPC datasets. It will allow data discovery by highlighting associated datasets to the user.

For example, the ESA-CCI ECV projects and the GCOS ECVs are controlled vocabularies; using SKOS these two controlled vocabularies can have their associative relationships defined. Using the workflow outlined in section 1 the domain expert creates spreadsheets in which relationships are defined as seen below:

GCOS-ECV URI	**GCOS has <relationship> CCI**	**ESA-CCI URI**	**GCOS has <relationship> CCI**
atmosphericUpperairWaterVapour	broadMatch	greenhouseGases	narrowMatch
atmosphericUpperairCloudProperties	broadMatch	cloud	narrowMatch
atmosphericCompositionCarbonDioxide	broadMatch	greenhouseGases	narrowMatch
atmosphericCompositionMethane	broadMatch	greenhouseGases	narrowMatch
atmosphericCompositionOzoneAndAerosol	broadMatch	ozone	narrowMatch
atmosphericCompositionOzoneAndAerosol	broadMatch	aerosol	narrowMatch
oceanicSurfaceSeaSurfaceTemperature	closeMatch	seaSurfaceTemperature	closeMatch

A more detailed graphical depiction of the CCI concept scheme is shown in Figure 2 A full description of this concept scheme can be found here (http://vm62.nubes.stfc.ac.uk/cci/cci-content/)

39 http://www.w3.org/RDF/

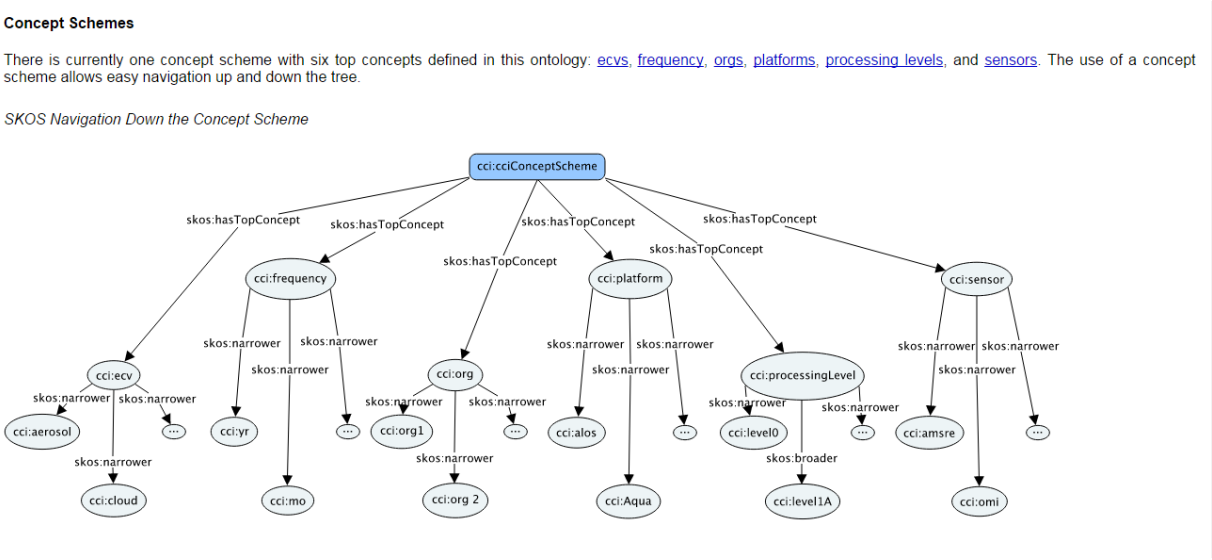

Figure 2: Graphical representation of the SKOS implementation for the CCI project.

Figure 2 illustrates how the controlled vocabularies of the CCI project: ECVs, frequency, orgs (organisations), platforms, processing levels, and sensors are constructed. Each concept is defined as a SKOS concept and is a member of one of these concept schemes. Each concept is also a sub-class of one of six (OWL) classes. In turn theses six classes are all members of an unordered collection, a Bag. The use of OWL alongside SKOS is discussed in http://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html. OWL allows for the imposition of a more formal structure which may then be used by reasoners when performing complex queries.

The next stage of the controlled vocabularies work is to progress with the inclusion of these within the NERC Vocabulary Server (NVS) and to continue to support the creation of any new and necessary controlled vocabularies, e.g. for the impact indicators. It is anticipated that the CMIP5 DRS controlled vocabularies will also be incorporated into the NVS in a similar way to the CCI shown above.

4. Catalogue (discovery) metadata

Catalogue metadata, also termed discovery metadata (Lawrence et al. 2009), are used to aid discovery of datasets. It is common for this type of metadata to be submitted to, or harvested into other catalogues and organisations to facilitate data discovery. It is intended that all CLIPC data be compliant with the ISO19115/19139 INSPIRE metadata standards (i.e. Gemini 2.2) so that the data is widely visible, and the catalogue records can be "consumed" by multiple organisations and systems.

4.1 Dataset level metadata specifications

For users to find data through the CLIPC Catalogue, we need consistent dataset level descriptions across data providers. Existing initiatives, including those producing contributing datasets, use ISO19115/19139 INSPIRE compliant metadata. The overarching concept for the catalogue is to harvest these metadata where they already exist (e.g. KNMI, STFC, SMHI (TBC)) as they will already generate ISO-compliant metadata records. New records will be

created where they do not exist and all records will be visible and searchable via the CLIPC Portal. As datasets are added, created or modified regular updates to the metadata catalogue will be needed and new metadata records will need to be validated against a schema, and a schematron where available (a schema (.xsd - XML Schema Definition) defines the format of the metadata, a schematron gives a rule based validation of the metadata XML content). Where there are issues, e.g. the metadata record does not validate, fixing the content will be a manual task. The exact implementation and processes to be adopted are being developed and tested in the architecture team. A number of similar metadata profiles are already in use by partners and in related activities, all of which are compliant with the ISO19115/19139 and INSPIRE protocols, but with some differences in e.g. the keyword vocabularies used.

The EU INSPIRE Directive mandates the collection of metadata for use in Europe. Implementing Rules define the requirements for metadata for discovery purposes. These are based on ISO 19115. The aim of UK GEMINI is to provide a core set of metadata elements for use in a UK national geospatial metadata service, that are compatible with the INSPIRE requirements for metadata. It does not preclude organisations recording additional metadata elements for their own internal business purposes.

The minimum fields that are required (and which are consistently included in all profiles considered) are listed below. Most terms are mandatory, except where listed as optional or conditional. These terms are fully described in the UK Gemini Specification for discovery metadata for geospatial resources, v2.1, August 2010[40]

- Title
- Alternative Title (optional)
- Dataset language (conditional – data resource contains textual information)
- Abstract
- Topic category
- Keyword
- Temporal extent
- Dataset reference date
- Lineage
- West bounding longitude
- East bounding longitude
- North bounding latitude
- South bounding latitude
- Extent (optional)
- Vertical extent information (optional)
- Spatial reference system
- Spatial resolution (conditional - for datasets and dataset series where a resolution distance can be specified)
- Resource locator (conditional when on-line resource available)
- Data format (optional)
- Responsible organization
- Frequency of update

40 https://www.agi.org.uk/about/resources/category/81-gemini?download=17:gemini-2-1

- Limitations on public access
- Use constraints
- Additional information source (optional)
- Unique resource identifier
- Resource type
- Conformity (conditional – required if claiming conformance to INSPIRE)
- Equivalent scale (optional)

This will enable a minimum conformant set of fields to be provided by all partners providing metadata records for the CLIPC catalogue, whilst allowing additional fields to be added for partners' internal use.

4.2 CLIPC processes for creation, maintenance and publication of catalogue records and metadata

One of the challenges of CLIPC is to coherently present catalogue metadata from data that use different metadata standards. Therefore the following roles and responsibilities are defined in order to allow metadata records from diverse sources to be searchable in the central CLIPC catalogue.

Role	Task
Data producer	Produce data; either creates a compliant XML file, or provides enough relevant information to another party who creates a compliant XML file. The XML file is given to the metadata curator. This process goes hand in hand with providing the data to the data curator/ data archive as organisation responsible for the dataset.
Data curator (e.g. STFC-CEDA, KNMI, SMHI)	Ensures catalogue record and curated dataset are consistent. Makes catalogue record available to catalogue operator.
Data publisher (e.g. STFC-CEDA, KNMI, SMHI)	Ensures catalogue record and curated dataset are consistent. Makes catalogue record available to catalogue operator.
Metadata creator (could be data producer, or data curator)	Ensures catalogue record and published dataset are consistent.
Metadata curator (e.g. STFC, KNMI)	Creates a compliant XML file.
Catalogue operator (MARIS)	Maintains the catalogue record, makes it available to the catalogue operator.

An example of what the downstream user would receive would be some GEMINI compliant information e.g.

```
<gmd:keyword>
   <gmx:Anchor xlink:href=""
xlink:title="sea_surface_temperature">CFSN0381</gmx:Anchor>
</gmd:keyword>
```

This then allows the Catalogue operator to pick up this information and using the links within to seamlessly link from the portal to the metadata for the entity.

5. Metadata and controlled vocabularies for quality control

Metadata and controlled vocabularies can also be used within the quality control elements of the CLIPC portal.

5.1 CORE-CLIMAX maturity matrix

The CORE-CLIMAX[41] System Maturity Matrix self-assessment of data is to be used for quality control. There are 6 major categories where assessments are made:

1. Software readiness
2. Metadata
3. User documentation
4. Uncertainty characterisation
5. Public access, feedback, and update
6. Usage

For each of these categories the assessment will assign a score from 1 to 6 that reflects the maturity of the climate data record with respect to a specific category. Maturity scores 1 and 2 establish research capability, i.e. all aspects of the CDR are still under development. Maturity scores 3 and 4 establish an initial operations capability; at this stage the climate data and associated material are available to the user community. Maturity scores 5 and 6 indicate full operations capability; at this stage the production of the CDR has been transitioned into operational environments, e.g., the whole processing process is under configuration management, fully automated and performance is monitored. A data provider will self-assess their data against the above and complete a maturity-matrix.

A controlled vocabulary of terms for each maturity level and in each category can be defined; e.g. usage, level 3 is "Benefits for research applications demonstrated". Any dataset that is included within the CLIPC data that has self-assessed against the standardised maturity-matrix can then have the controlled vocabulary terms associated with that climate data record.

5.2 Use of CHARMe for commentary metadata

CHARMe[42] is a system that allows users to view or create annotations to climate data records that describes how data has been produced, processed or assessed. Additional information that could be used to enrich the data could include (from CHARMe.org):

- Citations that reference a particular dataset;
- Results of assessments - reanalysis, quantitative error assessments;
- Provenance - processing algorithms and chain data source;
- External events that may affect the data - volcanic eruptions, El Nino, sensor failure;
- Supplementary dataset quality information - maturity, discontinuity, updates.

The aim is to make the CHARMe commentary metadata searchable in the portal in a similar way as the data to providing users with all information that maybe related to the data.

41 http://www.coreclimax.eu/
42 http://charme.org.uk/

Implementation of the CHARMe plug-in in the portal catalogue will allow users and data providers to add qualitative information about fitness for purpose of the various datasets, add links to relevant documentation, reports, etc. and flag up potential issues with the data. To meet early user requirements expressed in CLIPC, a notification functionality has been added to the CHARMe software, to allow users to "follow" datasets of interest, i.e. receive updates as annotations are added.

Controlled vocabularies can be incorporated within CHARMe, it has a facility to attach SKOS key words within a linked data framework and enables more flexible linking between CHARMe and other parts of the portal. This functionality is something that will be explored further within the CLIPC project.

6. Metadata and controlled vocabularies for uncertainty data

Uncertainty information for climate data records does not at present have a standardized format. A brief summary of some different approaches to dealing with different types of uncertainty within the framework of the CF conventions is given below.

The CF conventions[43] allow for a "*cell_method*" attribute to be attached to a variable, however these are limited to maximum, minimum, median, mid_range (average of max and min), mean, mode, standard deviation and variance. These do not cover all the requirements for describing data uncertainty for all of the different types of uncertainty estimates used within CLIPC data. In order to begin to reconcile some of the terminology used in uncertainty estimates a case study of the ESA Climate Change Initiative (CCI) ECVs has been performed is discussed in section 6.1.

UncertML[44] is a conceptual model and XML schema designed for encapsulating probabilistic uncertainties and aims to quantify uncertainties. The interoperable model can be used to describe uncertainty in a variety of ways including:

- **Samples**
- **Statistics**: including mean, variance, standard deviation and quantile
- **Probability distributions**: including marginal and joint distributions and mixture models.

The well-defined terms in UncertML can be used to help describe the uncertainty information provided in CLIPC, however it is not used in a formalized or consistent way.

As part of the CORDEX project SMHI developed a data reference syntax (DRS) for bias corrected data. This work provides a framework for consistently describing bias corrected data, including both the DRS and variable naming conventions. It is suggested that bias corrected data variable names should be appended with the term "adjust".

43 http://cfconventions.org
44 http://www.uncertml.org/

6.1 ESA-CCI Uncertainties case study

A subset of the ESA CCI project data held at CEDA was interrogated as a case study of how satellite climate data uncertainty metadata is currently handled. The data are provided in NetCDF format. The standard names, long names, comments and descriptions of uncertainty data from the Aerosol, Cloud, Fire, Greenhouse gases, Ocean colour, Ozone, Sea ice, Sea level, Sea surface temperature and Soil moisture CCIs were analysed. Data were available at processing levels 2, 3 and 4. A summary of this work is as follows:

1. **Inconsistent use of quality flags.**
   Some of the interrogated data included numerical quality flags, all contained a comment stating what the numerical value represented e.g. 0 = no data. There was consistency within each CCI on the quality flag usage but an inconsistency between different CCIs. The numerical value 0 was used to represent no data and best data, dependent on the CCI. This inconsistency could be confusing to data users.

   RECOMMENDATION: All CCIs should use the same quality flag descriptors. Discussions with the CCI data standards working group have indicated a desire to converge on using '0' as the good or best quality data flag.

   Alternatives to the status quo:
   - A binary quality flag of 0 = good and 1 = bad
   - A quality flag array:

Flag	**0**	**1**	**2**	**3**	**4**	**5**
Quality	best	acceptable	Low	worst	bad_data	no_data

   - There is also the option to add an attribute label with the CF conventions

2. **CF cell methods.**
   The CF-conventions allow for the attachment of some basic variable descriptors such as mean, mode, standard deviation and variance. Although these are available not all CCI projects make use of this functionality with the CF conventions.

   RECOMMENDATION: All CCIs should use available cell methods where appropriate, though this may not be appropriate for level 2 pixel products, in this case a sensible and fully descriptive long name is recommended.

3. **Ambiguous long names, multiple terms and inconsistency.**

   Ambiguous long names:
   Where no standard name for the uncertainty data exists the long name is used to describe the uncertainty. This is not done to a consistently quality throughout the CCI projects, there are multiple instances where the long name does not provide enough information for an end user to be able to understand exactly what the uncertainty given is actually describing. This is further complicated through the ambiguous way in which the term uncertainty is used. Uncertainty can be random, systematic or the sum of these (total). In

addition the uncertainty could be of for example, a measurement, the retrieval algorithm or additional post-processing.

Multiple terms:
It is often the case that there are multiple terms which could be used to describe the same phenomenon, some examples are provided

- Both standard deviation and root mean square are used, formally these are defined identically. This is also the random component of the uncertainty.
- Bias is also used is the same as the systematic uncertainty.

Inconsistency:
Terms used to describe uncertainty are not used consistently between different CCIs and sometimes not even with a single CCI.

RECOMMENDATION: To begin to reconcile some of the issues outlined above the CCI data standards working group[45] is to propose the following standard names to be included within the CF conventions. However the following could be used as fully descriptive long names until a standard naming convention is approved:

1. standard_uncertainty_of_<quantity>
2. standard_uncertainty_of_<quantity>_due_to_<uncertaintyType/adjective>_errors_from_<effect>
3. <standard_name>_error_correlation_of_<correlationType>

(NOTE: 1. is strictly the total standard uncertainty of the quantity from all uncertainty types and from all effects.)

The following terms apply:

standard uncertainty	one standard deviation of the estimated error distribution
quantity	measurand (quantity being measured)
adjective	random systematic correlated
uncertaintyType	subsampling instrument_noise time_adjustment (list is not exhaustive)
correlationType	length_scale time_scale scale_height (list not exhaustive)

There is scope here to develop controlled vocabularies on the adjectives, uncertainty type and the correlation type that could be used within the portal.

45 This information was provided by Caroline Poulsen (STFC). The work of Chris Merchant (University of Reading) and Jonathon Gregory (UK Met Office) are acknowledged.

4. Within some CCI projects additional auxiliary variables are provided in the files, it is not clear if these auxiliary variables are directly related to the quality of the primary observed variable.

   RECOMMENDATION: If an auxiliary variable is included and it directly relates to the quality of the primary observed variable it should be commented that this is the case using the "ancillary_variables" variable attribute, as described in the CF Convention

6.2 Plan for progress

A number of controlled vocabularies exist within UncertML, these could utilised throughout the CCI with the addition of terms such as those from the table of terms for describing uncertainty. This would allow for consistency across the CCIs and provide fully descriptive and well defined variable names within the CF framework. Proposal and implementation of relevant controlled vocabularies for describing uncertainty is within the scope of the CLIPC work.

There has been collaboration between STFC and the CCI data standards working group representative based at STFC. The outcome of this case study along with other material developed by the CCI working standards group has been distributed to the CCI projects for consultation on best practice. An Uncertainties Workshop will be held in Hamburg in February 2016 will facilitate discussions across various EU funded projects in developing a common approach to dealing with uncertainty information for future use. There will likely be a role for ISO 19157[46] which will be discussed in the Uncertainties Workshop meeting notes.

From the draft version of MS37: Methodology for a qualitative uncertainty assessment of climate impact indicators some additional controlled vocabularies that could be used within the CLIPC portal are apparent, for example:

- Source of uncertainty: Incomplete knowledge, Unpredictability
- Qualitative Confidence: High, medium to high, low to medium, low

This only gives an indication of some of the work with controlled vocabularies that is within the scope of CLIPC and is a work-in-progress.

7. Metadata and controlled vocabularies for impact indicators

Metadata for indicators (tier 1, 2, and 3 indicators) that are being produced in CLIPC, and made available through the CLIPC data services layer, need harmonizing and consolidating in the same way as described for all the "raw" datasets described above. Producers of indicators datasets have typically not had much exposure to the sorts of metadata standardization activities described above. It is anticipated that significant progress with this work will be made at a workshop (held in conjunction with the Infrastructure for the European Network of Earth System Modelling (IS-ENES) FP7 project) in February 2016 in Toulouse. Outcomes

[46] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32575

from this workshop should begin to develop metadata standards, indicate required controlled vocabularies and DRS specifications for impact indicators within the frameworks described in this document.

8. Conclusions/ Future plans

Much progress has been made in the development of controlled vocabularies and metadata for CLIPC data. A summary of progress presented in this document is detailed below:

Data reference Syntax

Data	Progress	Source
ESA CCI & GlobSnow	Complete	CCI/CLIPC
HadOBS	Complete	CLIPC
Global Model data	Complete	CMIP5
Regional Model data	Complete	Cordex
Bias Adjusted data	Complete	SHMI
Global Reanalysis data		
Regional Reanalysis data	Complete	UERRA
Impact Indicators	On-going	CLIPC

Controlled Vocabularies and SKOS

Many controlled vocabularies already exist and they are being incorporated in the CLIPC work, where appropriate vocabularies do not exist new ones have been created; Section 3.1.1 lists the new controlled vocabularies created to date. This work is on-going. It is anticipated that a number of new controlled vocabularies will be required particularly in relation to uncertainty and quality control information and the impact indicators. The controlled vocabularies have been incorporated into a SKOS framework to facilitate harmonised data searches and increased visibility of related data.

Uncertainty and Quality Control metadata

The uncertainty case study highlighted that the metadata for uncertainty information is not used consistently and that it is not always fully descriptive. There is a need for metadata standards development which can be adopted by data producers. It is intended that the self-assessment CORE-CLIMAX maturity matrix be used by the data providers to rate the maturity of their data. Including the matrix as a controlled vocabulary will allow the self-assessment score to be attached to their climate data records. Additionally inclusion of commentary metadata through CHARMe with SKOS keywords will enable more flexible linking between CHARMe and other parts of the portal.

Future plans

There are workshops planned for February 2016 on uncertainty data and impact indicators. This will help to resolve some of the outstanding issues and allow progress to be made.

Appendix A: File level standards developed in CLIPC

Some additional file level standards were required within CLIPC for ESA Glob Snow, detailed below.

A.1 ESA Glob Snow

The ESA Glob Snow standard follows the framework of the ESA CCI standard, but extends the some vocabularies:

Vocabulary	Term(s) added
Projects	GlobSnow
Processing Level	L3A
Variables	Snow Water Equivalent SWE Snow Water Equivalent SWE_max Snow Water Equivalent SWE_avg Snow Water Equivalent SWE_std_avg Snow Water Equivalent SWE_std

Appendix B: A Data reference syntax consolidation

It was noted that many terms (facets) of the data reference syntax (DRS) of the datasets identified in section 2 included terms that were named differently but were in fact describing the same thing. Therefore a DRS syntax consolidation exercise is currently underway. The progress is summarized in the Table: Data Reference Syntax consolidation, note that this work is on-going, not all datasets have been included and this table should be viewed as a work-in-progress. Where new datasets are being created (e.g. CCI and HadOBS) new facet names should not be used rather existing names should be re-used where possible.

In the Table: Data Reference Syntax consolidation where lines are highlighted as in the same colour, this indicates a dataset already in existence and where the facet names although different describe the same information. Part of this work will involve looking at the properties the facets are describing. This was discussed in MS19[47] under the Different Properties in Section 3.2 Categories of Categories. This described three properties that the facets may have view, focus and style, defined as follows:

View: There are many estimates of the state of the world, and, from simulations, many realisations of possible worlds. For example: model, experiment, ensemble; mission, analysis method.

Focus: 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). For example: realm, variable, ECV, location (and masking), time period.

[47] http://www.clipc.eu/media/clipc/org/documents/milestones/ms19_drsvocabs_april2015_final.pdf

Style: Data can be presented in different ways, varying sampling rate, file format, etc. For example: frequency, EO processing level, spatial mesh, data format.

It is anticipated that when this work is completed that a list of commonly used facet names for the three properties; view, focus and style will be made available. This should help data producers in selecting a data reference syntax quickly and ensure that redundant facet names are not produced inadvertently.

DRS FACET	SOURCE	DESCRIPTION	CMIP5		CORDEX		UERRA		HadOBS		CCI	
activity	CMIP5	Modelling activity, e.g. CMIP5	x	CMIP5	x		x		x	CLIPC	x	CCI
product	CMIP5	CMIP5 specifies output - but for hadobs this could be product-type	x	output					x	insitu		
institute	CMIP5	Institute	x		x		x		x	acron		
model	CMIP5	Model and its version	x									
experiment	CMIP5	Group of experiments e.g. in CMIP - rcp45	x				x					
Frequency	CMIP5	Temporal frequency of output, yr, mon, day, etc	x				x		x	hr/daily/mon		
modelling realm	CMIP5	CMIP5 controlled vocabulary of modelling realms e.g. atmos	x									
table	CMIP5	MIP lookup table	x						x	? include		
ensemble member	CMIP5	Ensemble member reference	x						x	only req for HadCRUT4		
version number	CMIP5	Version of publication-level dataset YYYYMMDD (ESGF)	x						x	vYYYYMMDD		
regional domain	CORDEX	A regional modelling domain			x							
RCMVersionID	CORDEX	Regional Climate Model version ID			x							
Bcname	SHMI	Bias correction methodology			x							
OBSname	SHMI	Reference for bias correction			x							
REFperiod	SHMI	bias correction reference period			x							
framework	MOHC	Dataset framework e.g. HadOBS							x	HadOBS		
collection	MOHC	Dataset eg HadISD, HadISDH, HadCRUT4							x	Dataset		
product_version	MOHC	Native version control							x	e.g v1.0.3.0p		
Domain	UERRA	Regional domain identifier					x					
RADrivingName	UERRA	Driving Model and its version					x					
RADrivingEnsmbleMember	UERRA	Driving ensemble member identifier					x					
RAModelName	UERRA	Name of regional analysis model					x					
RAEnsembleMember	UERRA	Ensmble member identifier					x					
variableName							x		x		x	
CCI project	CCI	The project producing the data: corresponds to ECV									x	
Processing level	CCI	EO processing level, using standardised CCI definitions.									x	
Data Type	CCI	Physical variable, e.g. Aerosol optical depth									x	
Product	CCI	dataset, sometimes referring to an algorithm									x	
Additional Seg.	CCI	that different post-processing choices (e.g. aggregation to monthly									x	
Indicative date	CCI										x	
Indicative time	CCI										x	
Data specification version	CCI	(ftp://podaac.jpl.nasa.gov/OceanTemperature/ghrsst/docs/GDS20r5.									x	
File version	CCI	should also result in an incremented file version number.									x	

Table: Data Reference Syntax consolidation

References

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:

Lawrence, B. N., Lowry, R., Miller, P., Snaith, H. and Woolfe, A., 2009: Information in environmental data grids, *Phil. Trans. R. Soc. A*, **367**, 1003-1014, doi:10.1098/rsta.2008.0237

GCOS (2010): Implementation plan for the global observing system for climate in support of the UNFCCC.