MS18 Extended specifications for climate data

Victoria Bennett, 29th April 2015, v1.0

CLIPC Task 5.2 Milestone: Metadata specification for observation, ESA Climate Change Initiative and re-analysis datasets

Task description (from DoW)

T5.2 - Metadata standards and services [Months: 3-25]
STFC, KNMI, SMHI, CNRS, Met Office
This task will extend metadata standards for data publication, quality indicators and uncertainty developed in a range other initiatives (CHARMe, CIM, HMA, CMIP5, CORDEX, EUROCORDEX). The ESGF metadata standards will be extended to encompass datasets identified in Task 5.1, including CCIs, extremes, observations and reanalysis products, in consultation with data providers and using experience already gained in extending ESGF for satellite observations (Obs4MIPs project) and regional modelling (CORDEX project). This work will leverage parallel effort being undertaken within the IS-ENES2 project. A conceptual model for quality control and uncertainty will be developed expanding on CHARMe. This model will include concepts for bias correction and uncertainty of climate model data as required by users and impact modellers. Controlled vocabularies for these standards will be developed utilising the vocabularies management services of Task 3.3.

Summary

This milestone document reports progress to April 2015 on extending metadata standards for data publication, building on work in other initiatives, and reusing experience gained elsewhere. In particular, the metadata profile for the CLIPC catalogue is discussed and decisions are summarised here.

Introduction

In CLIPC we need to consider different data types and formats from different communities, each with their own community-specific metadata specifications.

Here we aim to gather details of existing metadata standards for the datasets in scope for CLIPC (initially those in the inventory, D5.1) and propose a protocol for handling datasets that do not yet adhere to particular community metadata standards. We can't make all the data conform to one set of standards, but we need to pick a small number and try to make other datasets fit into one of those.

We also need to consider dataset level metadata for cataloguing the CLIPC data holdings to enable search and discovery in the CLIPC portal. There are a number of national and European projects already active that run high-profile catalogues, and comply with international standards (e.g. ISO19115/19139). For CLIPC, we need to find a solution that reuses existing work in this area but allows us to tailor the catalogue records as needed to meet the project's and the users' requirements.

File level metadata[1] specifications

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).

[1] Throughout this document "file level metadata" implies metadata, typically within a data file, describing the data within the individual file, whereas "dataset level metadata" implies a catalogue record, typically describing a collection of files/products

It is not practical to develop a new metadata standard for CLIPC and transform the heterogeneous data sources to meet one standard, however, where datasets are still being created we have an opportunity to guide the selection of metadata to agree with one of a number of existing and commonly used specifications.

The table below lists the specifications in use by the Dataset Inventory (D5.1)[2] datasets:

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	TBD (likely link to EUMETSAT system, not publish through CLIPC)
GlobSnow	Under development (WP5 activity); likely to be based on ESA CCI Data Standards
CORDEX	Based on CMIP5 specifications
SPECS	Based on CMIP5 specifications

In general, the model data follow specifications based on those developed in CMIP5, and the observation datasets will most likely follow specifications similar to those developed for ESA CCI. A number of datasets will also be produced within the CLIPC project (e.g. impact indicators and bias corrected model data). These should follow one of the metadata specifications mentioned in the table above.

A discussion of the some of issues with using a range of community metadata specifications, particularly in relation to vocabularies, is given in the CLIPC Report on Extended controlled Vocabularies[3].

Dataset level metadata specifications

CLIPC will offer a data product catalogue as discussed in the CLIPC architecture team report[4]. This will include metadata records describing the datasets within the CLIPC system. These metadata aim to enable the data to be better understood and used to good effect.

The catalogue will need to include records for all the datasets that can be accessed via the CLIPC portal. Some examples of what is in and out of scope are given in the table below:

Datasets in scope to be included in CLIPC product catalogue	**Datasets out of scope – not included in CLIPC product catalogue**
• Datasets published in CLIPC through WP5 (Data Access) • Bias corrected data produced in WP6 (Transforming climate data) • Processed Tier 1,2,3 datasets	• Data products created by users of the CLIPC toolbox • Datasets hosted and made available by institutions outside the CLIPC project (eg EUMETSAT CMSAF)

[2] http://www.clipc.eu/media/clipc/org/documents/deliverables/d5_1_climate%20dataset%20inventory.pdf
[3] http://www.clipc.eu/media/clipc/org/documents/milestones/ms19_drsvocabs_march2015_final.pdf
[4] http://www.clipc.eu/media/clipc/org/documents/other/clipc_at_v1_1_feb2015.pdf

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) already generate ISO-compliant metadata records), create new records where they do not exist, and make all the records visible and searchable via the CLIPC Portal.

As datasets are added, created, 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 will be agreed and progressed 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 following metadata profiles were analysed and compared, before making a decision on the profile to select for CLIPC.

- KNMI Data Centre
- UK Gemini 2
- MACC
- WMO Core
- SeaDataNet
- MyOcean

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.

While some of the profiles comprise additional topics (e.g. quicklooks), the categories of metadata fields that are addressed in all of the metadata profiles considered are:

1. Product Identification
2. Classification and keywords
3. Spatio-temporal extent
4. Organisations responsible for management and processing, points of contact
5. Product Access
6. Data Quality and Validity info

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[5].

- Title
- Alternative Title (optional)
- Dataset language (conditional – data resource contains textual information)

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

- 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 organisation
- Frequency of update
- 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)

All other metadata profiles considered (e.g. KNMI, WMO, SeaDataNet) include these fields but also a number of additional terms, which are of particular relevance or importance to their user community and/or data centre operations. It is therefore proposed for CLIPC to proceed with the UK GEMINI-2 profile initially. 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.

Some further discussion is still required to agree on the keywords to be adopted in the CLIPC catalogue (in the "Keyword" field).

The UK-GEMINI specification states : "Keyword values should if possible be taken from a list of standard subject categories, identified in the element „Originating controlled vocabulary". Possible vocabularies are the Integrated Public Sector Vocabulary (IPSV)10 from the esd-toolkit, which should be used by public sector bodies, or the General Environmental Multi-Lingual Thesaurus (GEMET11), which should be used for INSPIRE conformance"

The intention is to mandate both a general keyword vocabulary (e.g. GEMET, mandatory in INSPIRE) to support cross domain data discovery, and additional specific keywords specific for our own convenience (as data curators, publishers and users), e.g. identify the variables in the files with parameters from CF standard name table, or use other vocabularies if appropriate. This approach is analogous to that adopted by SeaDataNet and MyOcean.

Another field which can be used in different ways is "Resource locator". The UK GEMINI specification states "Specify a valid URL to a dataset, series or service. If no direct link is available, a link to a point of contact where more information is available may be given"

In some cases, this field contains a long list of resources, such as data download, visualisation, documentation. In other cases it is only the data download location. CLIPC will form a recommendation for what should be included here for the CLIPC catalogue.

CLIPC processes for creation, maintenance and publication of catalogue records

The following roles and responsibilities are defined in order to allow metadata records from diverse sources to be searchable in the central CLIPC catalogue.

Data producer: produces data, and 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 published dataset are consistent.

Metadata creator (could be data producer, or data curator): creates a compliant XML file

Metadata curator (e.g. STFC, KNMI) : maintains the catalogue record, makes it available to the catalogue operator

Catalogue operator (MARIS): periodically harvests metadata records from metadata curators and enables user search functionality