CLIPC Milestone 20: Initial Data Publication

Victoria Bennett, June 2015

Introduction

One of the aims of WP5 is to build a data services layer to the different datasets, located at various data centres and institutes, which act as a back-end for the integration services developed in WP4. The data services layer will build on the distributed archive infrastructure of the Earth System Grid Federation (ESGF), which already provides support for CMIP5, CORDEX, and the obs4MIPs Earth observation data. Datasets identified by the project will be prepared for publishing into the system by reprocessing into standard formats, developing standard metadata and extracting from tape archive.

Deliverable D5.1 : Climate Dataset Inventory[1] provides an initial inventory of climate science datasets for inclusion in the CLIPC catalogue and use by the toolkit. The architectural approach to the data services layer is detailed in deliverable D3.1 : Conceptual design of the CLIPC Portal[2] , and the Architecture team report[3] .

This milestone concerns the publication of initial inventory datasets through the data services layer.

What is CLIPC Data Publication?

Data will be made visible, accessible to the CLIPC data services through publication from an ESGF data node. If the data in question are not already hosted at an institution with an ESGF node, they need to be transferred to one of these locations (within CLIPC, STFC-BADC and LIU have ESGF nodes).

Some work is typically required to prepare the data for ESGF publication. This includes agreeing a DRS (Data Reference Syntax), format, metadata, vocabularies directory structure and file-naming convention, then converting the data to comply with the agreed structures. Further details of the metadata specifications and vocabulary issues can be found in other CLIPC documents. (e.g. M18: Extended Specifications for Climate Data[4] and M19: Extended Controlled Vocabularies[5])

The next steps involve compliance checking, then publishing the data, using a mixture of bespoke and generic software developed in the framework of the Earth System Grid Federation.

In ESGF, datasets are associated with “projects”, e.g. CMIP5, OBS4MIPs. This project is defined in the DRS and appears as a facet in the search interface, helping users find the data they need (see figure 1). Some of the datasets included in CLIPC services are already associated with existing projects (eg SPECS, CORDEX). Datasets that are not yet associated with existing activities will be identified as project ”CLIPC”.

---

1 http://www.clipc.eu/media/clipc/org/documents/deliverables/d5_1_climate%20dataset%20inventory.pdf

2 http://www.clipc.eu/media/clipc/org/documents/deliverables/d3%201_clipc_concept_design_portal.pdf

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

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

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

Figure 1: ESGF search interface showing "project" facets already included in ESGF published data ; new project "CLIPC" has been added

The ESGF workflow is not optimised for handling heterogeneous datasets: whilst climate simulation data can be made to be reasonably uniform in their layout and format (much preceding work has gone on in this area for CMIP5), the addition of satellite and in situ observations, reanalyses, etc. adds complexity which can require considerable effort and problem solving to ensure a robust and reusable technical solution. The CLIPC project partners, as well as collaborators in the ESGF, have been working on these topics during the past months.

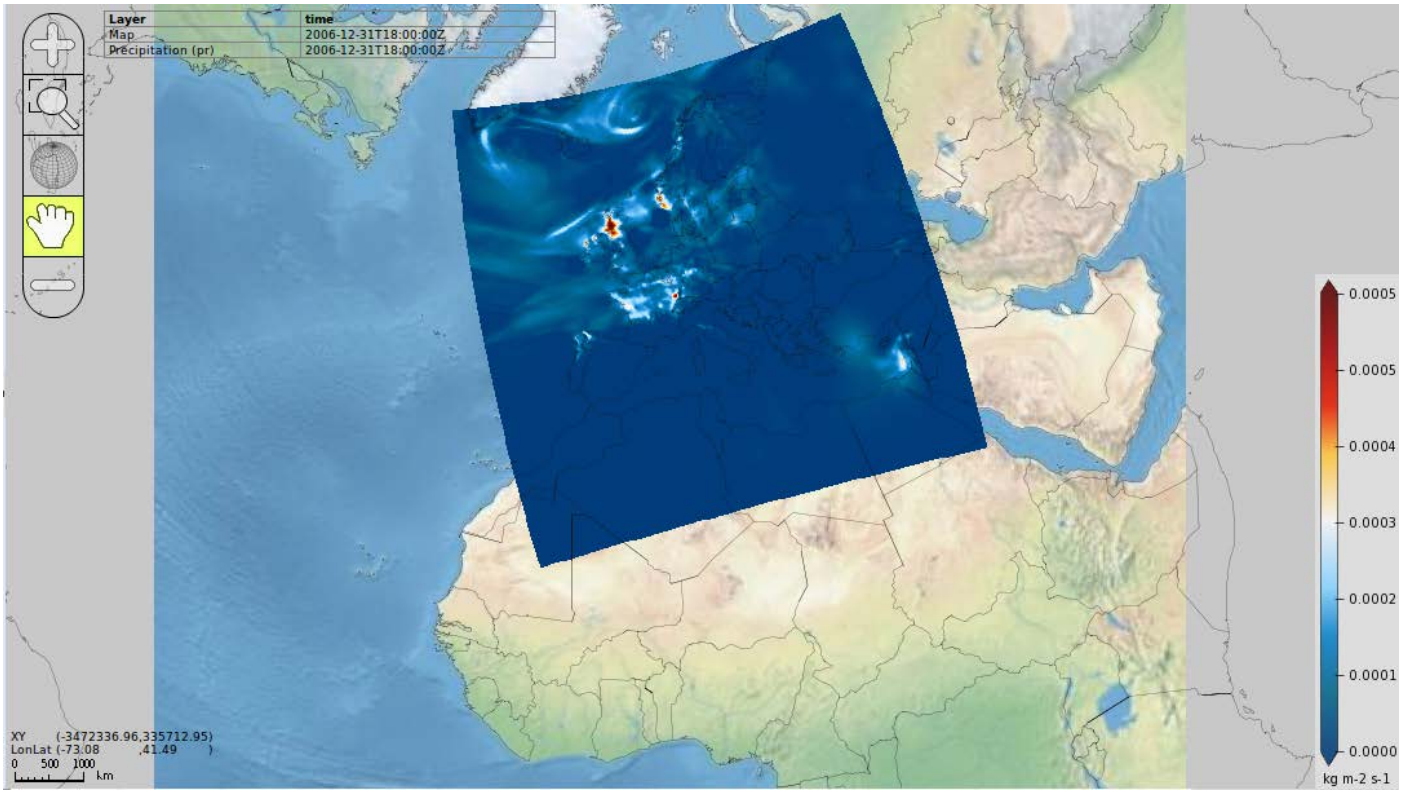

Figure 2: Visualisation of the SMHI-MESAN regional re-analysis data in the climate4impact portal, exploiting dynamic data access from the ESGF service at SMHI.

Key stages

Meta-data specifications

Clean visualisation of data in generic tools such as the climate4impact visualisation illustrated in Figure 2 depends on well structured meta-data. As the ESGF infrastructure is new and evolving, this requirement is unfortunately not well quantified at present. Meta-data specifications for ESA-CCI data were provided the ESA CCI programme: systematic checking by CLIPC led to some clarifications in a revised version. New meta-data specifications have been completed for regional re-analysis (SMHI) and are in development for surface observations (STFC, UKMO). The complexity of the data structures associated with observations, with a greater range of inter-related variables and associated quality control flags, need to be encoded carefully in the CF Conventions framework.

Compliance checking of data

The meta-data specification may contain specific requirements for dozens of items which need to be specified in the header of every data file. Reliability is enhanced by running compliance checks prior to publication of the data. Compliance checks have been run on two ESA CCI datasets and the ESA Glob Snow data.

Publication in ESGF

The ESGF distributed archive provides a unique facility to distribute data globally through an integrated system with a single global index. The ESGF archive has been designed primarily for climate projections. Some Earth Observation data has been published in ESGF through the obs4mips project, though in a form which is constrained to shadow the formats of the climate projections. The challenge in CLIPC is to enable use of ESGF for a wider range of data types without constraining the structure of the data.

Progress

This section provides an overview of progress with initial data publication into ESGF, for integration into the CLIPC portal.

Some of these datasets will appear in the "CLIPC" project in ESGF, others appear via other projects, but all will be accessible by the CLIPC data access and services layers.

Dataset *(and source location)*	**How**	**Where will the data be held**	**Status**	**ESGF "Project"**
EURO4M HIRLAM/3D-VAR reanalysis *(SMHI)*	Publish through SMHI/LIU ESGF node.	LIU/SMHI	Awaits implementation of the SODA tape-to-ESGF interface (CLIPC Deliverable D5.4, Month 30)	CLIPC
EURO4M MESAN 2-d downscaled surface reanalysis *(SMHI)*	The four key variables are published through SMHI/LIU ESGF node	LIU/SMHI	Done	CLIPC
ECA&D eOBS *(KNMI)*	STFC-CEDA take a copy and publish through BADC ESGF node. Need to transfer the data, possibly transform it, publish it. Make sure the vocabs work for this	KNMI.	Not started	CLIPC
CCI SST and Ocean Colour data (Level 3 only) *(ESA CCI projects)*	STFC-CEDA have a copy, need to publish through ESGF (vocabs to be agreed)	CEDA	Specifications agreed. Compliance checks completed. Ready for publication. Test data (subset) will be published June 2015.	CLIPC
CCI – other products *(ESA CCI projects)*	Will be published to ESGF via ESA Open Data Portal Project	CEDA	Starting May 2015	ESACCI (TBC), and subsets via Obs4MIPs
Globsnow *(FMI)*	Data to be transferred to STFC-CEDA from FMI for publication	CEDA	Transferring data for checking and publication	CLIPC
HadOBS *(Met Office)*	Transfer data to STFC-CEDA from Met Office for publication	CEDA	Agreed priority subsets for publication. Discussing specifications before transformation, transfer then publication	CLIPC

EUMETSAT CMSAF *(EUMETSAT)*	TBD	EUMETSAT	Doesn't make sense to publish to ESGF, but could link directly from CLIPC Portal.[i]	-
CORDEX	Already published in ESGF	Multiple[6]	done	CORDEX
SPECS	Being published in ESGF through SPECS project	STFC, LIU (SMHI)	underway	SPECS
CMIP5	Already published in ESGF	Multiple[7]	done	CMIP5

[i] This approach means that the data can be found via the CLIPC Catalogue, but they will not be available through the CLIPC data access and service layers for visualisation, processing etc

6 : As of June 2015: Linköping Uni. (with SMHI), DKRZ, DMI, Uni. Cantabria, Indian Institute of Tropical Meteorology, IPSL, Metéo-France, STFC

7 : As of June 2015: DKRZ, IPSL, STFC, UCAR, LLNL, Met No, NCI (Australia), CINES (France), NASA, NOAA, ICHEC (Ireland), NERSC (USA), LASG (China), CMA (China), BNU (China), FIO (China), CMCC, CCCMA (Canada), DIAS (Japan),