Integrating ES-DOC with the ESG Publisher service

Alan Iwi* (alan.iwi@stfc.ac.uk), **David Hassell**† (david.hassell@ncas.ac.uk), **Mark Greenslade**‡ (momipsl@ipsl.jussieu.fr) & **Ag Stephens*** (ag.stephens@stfc.ac.uk).
*STFC Centre for Environmental Data Analysis (CEDA), UK. †NCAS/University of Reading, UK. ‡Institut Pierre Simon Laplace (IPSL), France.

Climate modelling is a highly sophisticated process that generates petabytes of complex simulation data. In order to support the discovery and exploitation of these outputs, the international scientific community has developed tools to manage both the data and the detailed contextual metadata that describes it. The ESG Publisher, part of ESGF, scans netCDF metadata to enable data discovery, metadata interrogation, data download and sub-setting at both the *file* and *dataset* level.

The cdf2cim tool, part of ES-DOC, scans the same files to capture information that represents each s*imulation*. This connects the data itself to higher level concepts, such as *model* and *experiment*, that have been described through other means. The integration of cdf2cim and the ESG Publisher enables automated collection and cataloguing of simulation metadata as a key interface between the ES-DOC and ESGF worlds.

ES-DOC and cdf2cim

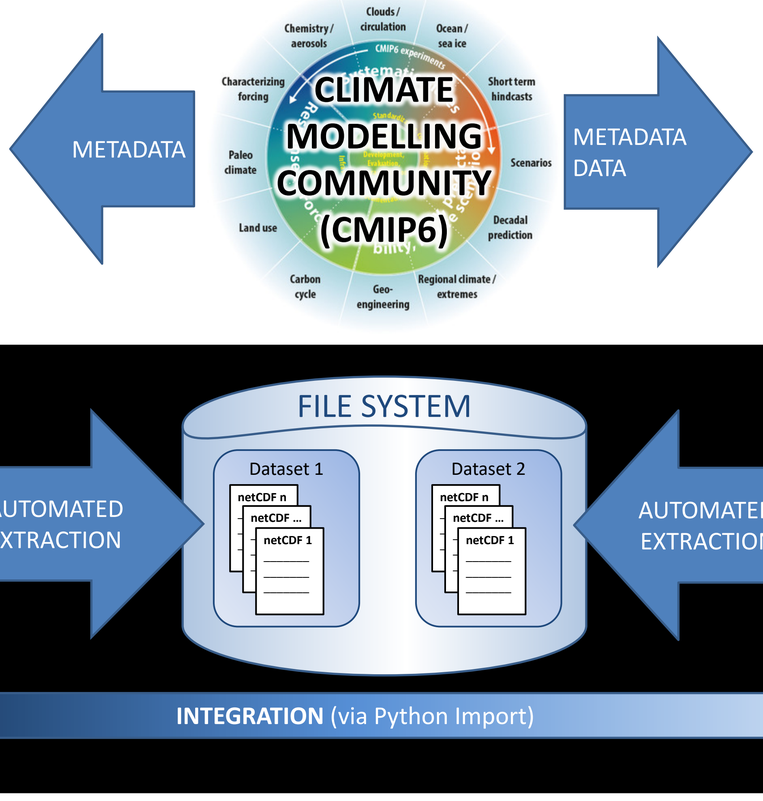
The Earth System Documentation (ES-DOC) ecosystem is able to capture, describe and disseminate essential information about climate modelling activities. Within **CMIP6**, scientists are describing their *models* and *experiments* in detail using a rich semantic model known as CIM2 (Common Information Model 2). ES-DOC also requires information about *ensemble* runs and each individual *simulation*.

A simulation document is an independent entity, but all simulations are considered to be part of an ensemble. The ES-DOC tools (https://es-doc.org) allow users to select and compare simulations. A CIM2 simulation document describes *a single integration by a particular model and why the integration was carried out*. Time span, ensemble and parent simulation properties are also described.

Automated metadata extraction

For CMIP5, simulations had to be documented manually. As a result, many were undocumented and there were often errors in records that were produced. The extensive global metadata in CMIP6 netCDF data files provides sufficient information to allow the simulation content to be extracted by scanning the files directly. Automation has the potential to ensure that everything is completely documented and based on the actual archived data files.

A command-line tool and Python library, **cdf2cim**, has been developed to manage the file scanning, serialisation to JSON, and upload to the ES-DOC server. cdf2cim is packaged so that it can be imported by the ESG Publisher.



ESGF and the ESG Publisher

The **Earth System Grid Federation** (**ESGF**) is an international collaboration for the software that powers most global climate change research, notably assessments by the Intergovernmental Panel on Climate Change (IPCC). Petabytes of high-profile climate simulations are archived and replicated across the globe. ESGF provides tools and interfaces for data management, discovery, search, browse, extract, subset and download to a diverse range of sectors interested in climate change and impacts.

Automated metadata extraction

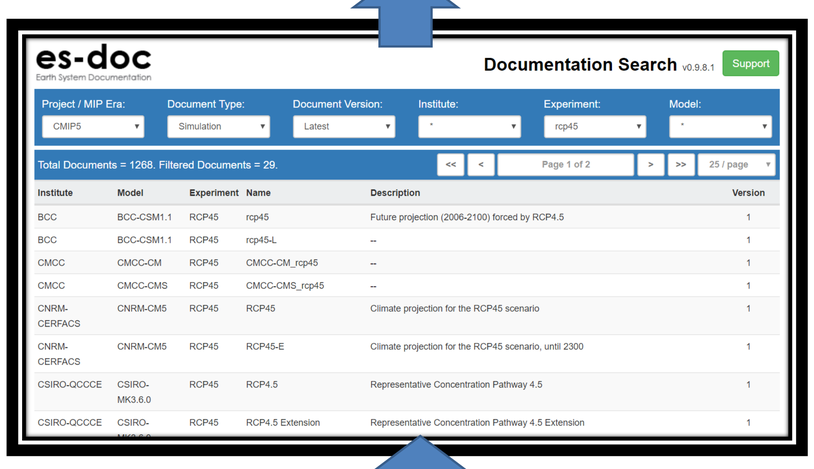
The **ESG Publisher** captures information from the netCDF data files to generate aggregations and metadata summaries suitable for publishing to various sources, including the THREDDS data server and the ESGF Search system. The metadata extracted during this process is used to underpin the user interfaces that allow interrogation and access at both the file and dataset levels.

Integrating the ESG Publisher and cdf2cim

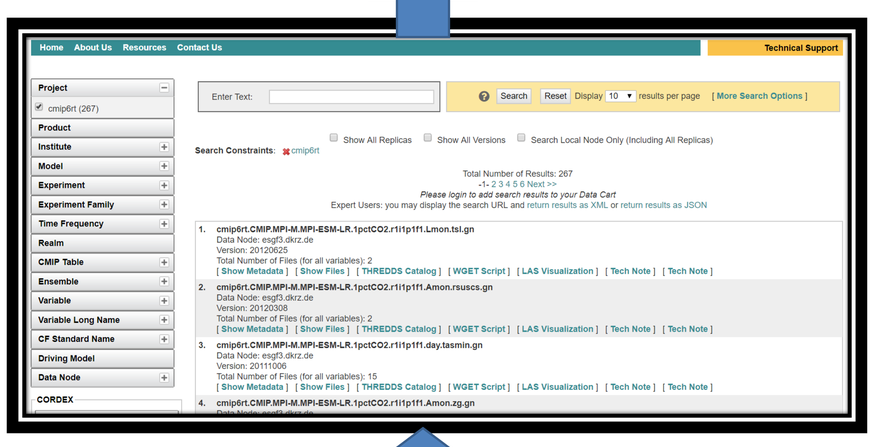
Since all CMIP6 data (within ESGF) will pass through the Publisher and every netCDF file is read, it is logical to extend the Publisher interface to extract CIM2 content at this stage in the workflow. This means that the publication process will include a second scan of the files by importing the cdf2cim package directly. For CMIP6, the default behaviour will be to call cdf2cim, via the *CMIP6 Handler*, but the integration will also allow it to be specified at the command-line, via a number of new command-line arguments.

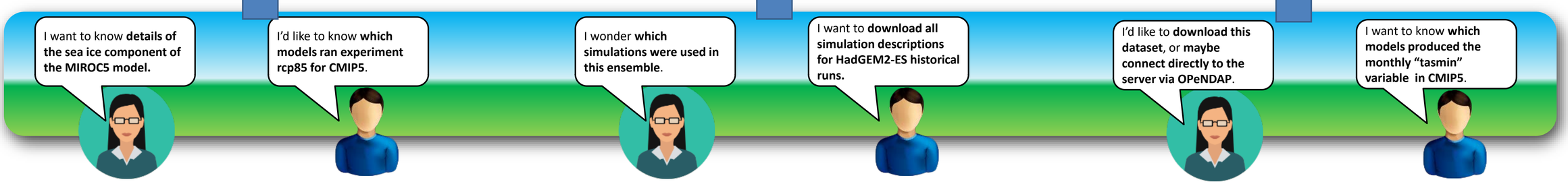
Manually generated metadata ⟷ **Auto-generated metadata** ⟷ **Data**

MODEL ENSEMBLE DATASET
EXPERIMENT SIMULATION FILE




```
# Import library.
import pyesdoc

# Import cim v2 schema.
import pyesdoc.ontologies.cim.v2 as cim

# Search remote archive: CMIP6 experiments.
experiments = pyesdoc.search('cmip6', 'experiment')

# Search result is iterable.
for i in experiments:
    print i.name

# Load a document from remote archive.
e = experiments.load_document('1pctCO2')
```





A GUIDE FOR DATA NODE MANAGERS

cdf2cim INSTALLATION

The standard installation of the ESGF Data Node includes installation of both the `esg-publisher` and the `cdf2cim` library. If you are installing the `esg-publisher` in stand-alone mode then the `cdf2cim` library will be installed when you resolve the dependencies using an appropriate package manager (such as `pip`). The `cdf2cim` dependencies are described in the requirements file at:

https://github.com/ES-DOC/esdoc-cdf2cim/blob/master/requirements.txt

References

ESG Publisher: https://esgf.github.io/esg-publisher/
CIM2 Overview: https://es-doc.org/cim
CDF2CIM GitHub: https://github.com/ES-DOC/esdoc-cdf2cim
CDF2CIM: https://es-doc.org/utility-library-cdf2cim

ES-DOC SERVICE: INTERACTIONS AND INSTRUCTIONS

Each simulation record generated from a collection of netCDF files is stored in a local JSON file in the directory: `$HOME/.esdoc/cdf2cim/scanned/`
When called by the `esg-publisher`, `cdf2cim` will attempt to upload the contents of any JSON files found in that directory to the ES-DOC server. After successful upload the JSON files will be moved to: `$HOME/.esdoc/cdf2cim/published/`

The size of each simulation record is very small (O(1kB)) so this should not impact on the Data Node or Publisher installation/service.

ES-DOC SERVICE: AUTHENTICATION

The ES-DOC web service will only accept content from authorised users. In order to gain access, each Data Node Manager will need to:

1. Request access to the appropriate GitHub Team.
2. Generate an authorisation token using GitHub and set the required environment variables to use GitHub credentials.

A test URL is provided to allow Node Managers to that test their credentials are valid.

USING cdf2cim WITH THE ESG PUBLISHER

The `cdf2cim` library is enabled using "`create_cim=True`" setting in the `esg.<project>.ini` file. This is set by default for **CMIP6**.

If you wish to include the call to `cdf2cim` when running for other projects, then you will need to add the line above, or you can explicitly call it with one of the following arguments to the `esgpublish` command:

`--create-cim:` Can only be used in conjunction with `--map`. Calls `cdf2cim` as an initial step before publication to the database, to create CIM records.

`--create-cim-only:` As `--create-cim`, but does not actually publish the dataset.

Additional command-line arguments that may be useful are: "`--no-create-cim`" (do not call `cd2fcim`) and "`--verbose-cim-errors`" (in the event of failure to publish CIM documents, display server errors).