Formal documentation for CMIP6

by the ES-DOC team

Version history:

- *Draft v. 0.1 – January 2015 Bryan Lawrence.*
- *Draft v. 0.2 – 30 Nov. 2016 Eric Guilyardi, sent to es-doc group for first round of contributions*
- *Draft v 0.3 – 12 Dec. 2016 Eric Guilyardi, including contributions from Allyn, Mark E, Sylvia, David*
- *Draft v 0.4 - 11 Jan 2017 Eric Guilyardi. including contribution from Mark G, David, Charlotte*
- *Version 1.0 - 18 Jan 2017 Eric Guilyardi. including comments from Allyn T. and Joseph Barsugli - version sent to WIP for review*
- *Version 1.1 - 05 Apr. 2017 Eric Guilyardi, including comments from WIP and new exec summary - version sent for community review*
- *Version 1.2 - 03 May 2017 Eric Guilyardi, following comments from Paul Durack*

Executive summary

This document provides an overview of the process for documenting CMIP6. It outlines the items to be documented, such as experiments, models, and simulations used to generate the CMIP6 datasets. It lists the key properties and features of these documents, the evolution from CMIP5 including lessons learned, the underlying tools and workflows as well as what modelling groups should expect and how they should engage with the documentation of their contribution to CMIP6.

The general approach has been to simplify and streamline the process as much as possible so as to ease the work of the modelling groups and adapt to their internal timelines. Most documents will be created either automatically or by internal ES-DOC effort, with minimal input from the modelling groups when and as needed. Documenting the model remains the main effort but, unlike for CMIP5, groups will both have several creation tools available and will be able to start from their existing CMIP5 model description (as available on https://search.es-doc.org/). These new CMIP6 documents can be created independently and in the order the groups wish. Their connection and access via the further-info-URL global attribute of the netCDF files will be dealt with directly by ES-DOC.

The documentation builds on the Common Information Model (CIM) standard build for CMIP5 and significantly updated for CMIP6. The document creation and update workflows and the different tools and processes available to groups are also detailed in this document. The new role of an institute-appointed ES-DOC officer acts as the main liaison between the ES-DOC team and the institution throughout CMIP6.

1. Introduction

Model intercomparison projects (MIPs) are essentially about designing experiments with specific requirements. They involve groups running simulations which conform to the requirements and which produce data that can be compared with both other simulation data and observations. Interpreting the data and the comparisons can depend on understanding the details of the models used and how they conform to the experimental requirements.
Historically this process has depended on well-constructed data specifications (e.g. Taylor and Doutriaux, 2010) so that data can be compared easily and on using information about the

models from a range of sources - from personal contacts, to websites, internal documents, and published papers. However, the latter rarely fully document the current state of models, generally concentrating solely on key new features. This lack of a single source of comprehensive information, coupled with significant complexity, means that model data users have been required to interpolate information both from the published record and a "form of institutional wisdom". Going beyond this is important not only for scientific understanding of model differences but, under the increased scrutiny of society, it is also demanded of a science that purports to be mature, credible and open to non-experts (Guilyardi et al. BAMS 2013).

This problem was formally addressed for CMIP5 by two parallel activities: the CMIP team developed a set of standards for the data output which ensured that data files held comprehensive metadata about the origin of the data, while at the same time, the Metafor and Curator project teams developed a "Common Information Model" (CIM) to encode more detailed information about the simulation workflow, and in particular, the models used and how they were configured for particular experiments. CMIP documentation has the ambition to extend what appears in the course of usual scientific publication and provides means to keep an up-to-date quality controlled documentation for the various types of users of CMIP datasets.

Much of the software required to capture the CMIP5 simulation workflow arrived late in the process, and the software to exploit that information was even later. As a consequence the system was inadequately tested and documented resulting in user confusion. Quality control for the metadata acquired came even later. Nonetheless, some excellent information was captured, and is available at http://search.es-doc.org, for community use.

Although we now have appropriate software for creating and manipulating simulation documentation, we can still learn significant lessons from CMIP5, including the necessity to clearly describe what the documentation is for, how it is being collected, and how it is to be exploited, before the process begins. This paper for the WGCM WIP and modelling groups describes that process for CMIP6 (Eyring et al. 2016). It first reviews the key aspects of the model intercomparison workflow to document, then presents the basic concepts of the new version of the CIM introduced to address the issues identified during CMIP5, and goes on to detail the processes and workflows used to build the CMIP6 CIM documentation (from a modelling group perspective) and its use (from a data user perspective).

This paper and its electronic version (https://es-doc.org/cmip6) will be updated regularly until early June 2017 following beta testing from a few modelling centers (phase 1: GFDL, Met Office, IPSL; phase 2: CCCMA, IITM and IAP).

In this context, the CMIP6 documentation effort explicitly addresses timeliness, structure, and process. In summary, we will:

1. Have the complete ecosystem ready and working and documented by early June 2017, some of it well before then, including full beta testing beforehand. We will include a complete end-to-end system using input data from a few pilot groups and the prototype tooling.
2. Make a clearer distinction between the different parts of the system, including ownership and governance, to ease creation of documents and the use of the suite of ES-DOC (and other) tools.

3. Allow a wider range of different individuals to take responsibility for information production and quality control of CIM documents, which will both speed up the process and make it easier.
4. Ease the load on modelling groups, namely by automation of the simulation metadata production, reporting by exception where possible, and providing alternative creation tools for model documentation (including scripting).

With these improvements, we believe it will be possible for some key information artefacts to go through peer review. In particular, we propose that

5. Both the final formal experiment and model descriptions reports appear in GMD.

If the WIP wishes, it is possible that the experiment descriptions could appear in the CMIP6 GMD special issue, allowing these too to both be peer reviewed and available as machine readable input (i.e. JSON) to institutional workflows.

2. The Modelling Workflow

In what follows we define the following terms:

- An experiment is an activity aimed at addressing a specific scientific problem, whether it is the simulation of one or more specific phenomena, the understanding of a specific real world process, or the understanding of the numerical behaviour of one or more models under idealised or real world conditions. We formally describe such an experiment by means of the **NumericalExperiment** class which describes the experimental aim, and is composed of a set of **NumericalRequirement** instances which should be met to address that experimental aim. These include any spatio-temporal constraints (what domain is simulated, and for how long), forcing constraints (e.g. whether a historical or future scenario is used for anthropogenic emissions of radiatively important gases), etc.
- A **Simulation** is a run of a configured **Model** which conforms to the **NumericalRequirements**, runs on a **Platform** and produces output **Datasets**.
- An **Ensemble** is a set of **Simulations**

Bold faced terms have special significance as they are represented in the CIM as "documents", of which more later. The complete workflow in the context of a model intercomparison project is depicted in Figure 1, where we see the concept of a "configured" model elaborated: most models have a number of possible configurations (possible resolutions, process sets, parameters, input datasets etc.), but any given simulation can only use one configuration.

There are six phases of this workflow that need addressing for complete documentation:

1. Describing the experiments and their requirements (e.g. MIPs) in such a way that the modelling community can execute their simulations with confidence that they have configured their own workflows and models in support of a particular experiment.
2. Documenting the actual model configurations used and what scientific processes those models encapsulated.
3. Acquiring information about which simulations were run and what data is available.

4. Linking that simulation information to important information about how those simulations conformed to the various experimental requirements (including information about how the simulations might have varied along ensemble axes, i.e. ripf attributes).
5. Recording the amount and type of computing used to deliver the experiments (to support assessments of future computing requirements), and
6. Documenting the data products themselves.

The documentation of the data products themselves is out of scope for ES-DOC (it is described in other WIP documents), but in the remainder of this paper we address the other five phases of documentation starting with the standard needed to encapsulate it: the CIM.

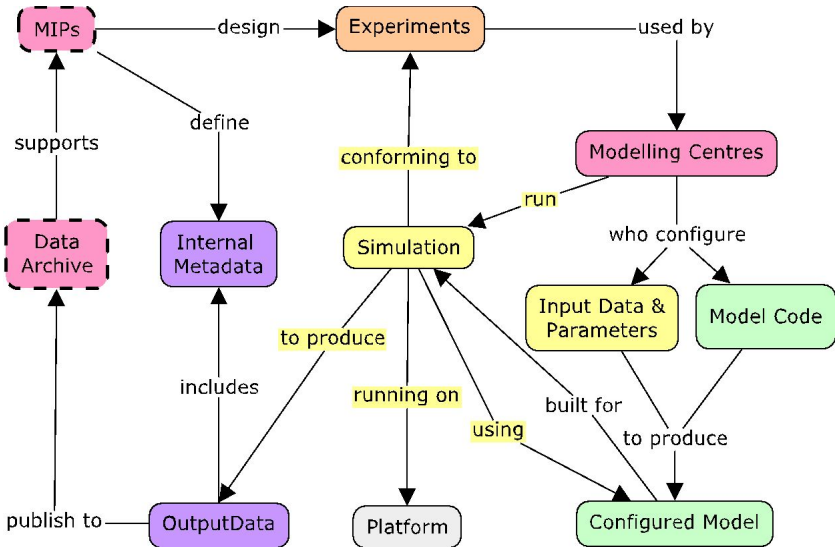

Figure 1: Simplified schematic of the modelling workflow associated with a model intercomparison project (MIP). MIP's design experiments which are used by modelling centres to configure their models so that they can run simulations which conform to the experimental requirements. The simulations produce output which have internal metadata and which are loaded into data archives. The different colours are intended to indicate the different domains of information: purple indicates information in and about data; yellow indicates key information about simulations, green about models, pink about modelling centers, orange about experiments. Dashed pink are coordinated from the WIP level and used by ES-DOC. Further details are discussed in the text. Note that "configured model" is the same as "model" in the following CMIP6 use.

3. Capturing the concepts in the CIM

The "Common Information Model" (CIM, Lawrence et al. (2012)) introduced the concept of CIM documents, that is, metadata artefacts created to describe different parts of the simulation workflow as described in Fig. 1. The revised version for CMIP6 is CIM 2.0 which is described in more details in Appendix I. The design criteria for CMIP6 documentation include making the infrastructure less intimidating than for CMIP5, and to streamline production by reducing complexity and duplication and increasing automation. By design there is no necessity for documents to be created in any particular order, and the connections are made by controlling the terms used to name entities. The canonical representation of the

CIM 2.0 schema (i.e. formal representation) is a simple pythonic format which simplifies downstream tooling chains.

4. Using the CIM for CMIP6

From the conceptual model of the CIM, the CMIP6 formal documentation is organised in the following documents, listed on the right hand side of Figure 2. The main documents listed in **Fig.2**, their creation and their handling, including what is required from modelling groups are described below.

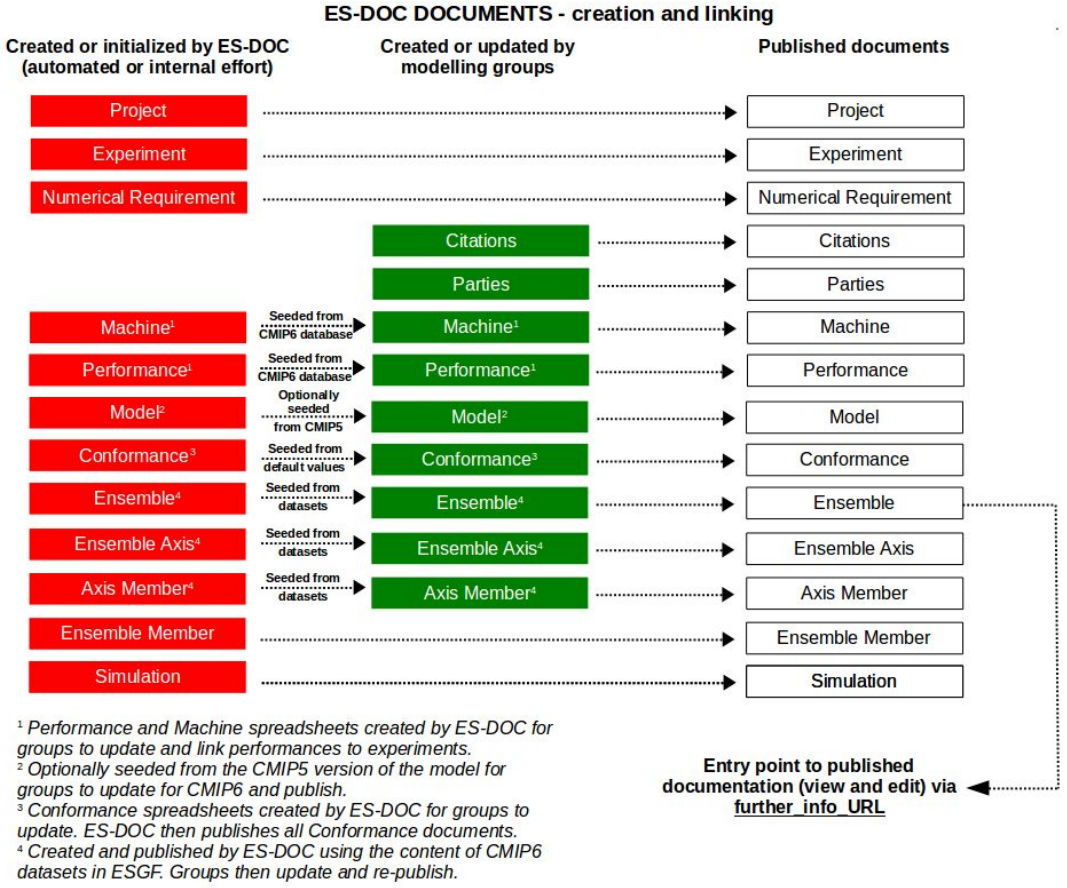

Figure 2. Overview of ES-DOC documents and creation workflow for CMIP6. The right hand side lists all documents and their relationships: red arrows denote a link automated by the ES-DOC software and the blue arrow denotes the only connection that has to be made by the modelling groups' ES-DOC officer. The further_info_URL points to the ensemble document and is the entry point to the full documentation. The documents listed on the left are generated by ES-DOC (either automatically or via internal effort). The central column lists the documents either created by modelling groups (Citations, Parties, Machine and Model) or updated toward their final version.

4.1 Experiment and numerical requirements

There are two key requirements for the documentation of experiments:

1. For modelers: to provide unambiguous, easily understood information about how to execute a compliant simulation, and
2. For data users: to understand when and how data from these experiments can be used.

In practice these both come down to good documentation of the "numerical requirements" of the experiments, that is, the constraints which modellers need to meet to run these simulations. The set of constraints supported by CIM2.0 is shown in **Fig. 3**.

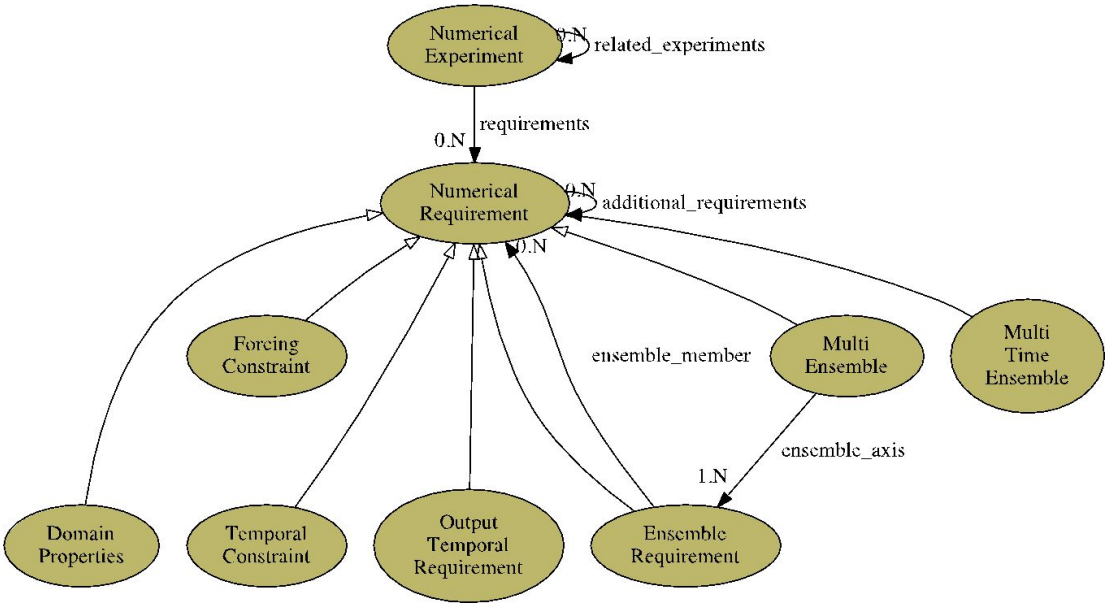

Figure 3: Numerical experiments and their requirements: there are many sub-classes of numerical requirements, ranging from constraints on the domain properties expected (e.g. global, Africa domain etc.) to multi-time ensemble constraints which, for example, describe an ensemble which requires a set of particular start dates. In this figure, arrows with hollow triangles point from sub-classes to super-classes, i.e. a forcing constraint is a numerical requirement. Arrows with solid heads point in the directions of associations, i.e. a multi-ensemble is associated with at least one and potentially many (N) ensemble requirements, similarly, numerical experiments may be associated with any number of other numerical experiments..

The descriptions of experiments in CIM2.0 for CMIP6 MIPS is now publicly available on http://search.es-doc.org/. A three stage review procedures has been organised:

1. All experiment descriptions to be discussed with the experiment originators,
2. A sub-panel of the WIP to be convened to address how the experiments should be presented, and
3. The collected experiments should be described and submitted to GMD, with the complete digital descriptions linked and available for download in multiple formats.

No action beyond reviewing is expected from the modelling groups for these documents.

4.2 Model

This is most demanding part of documentation for modeling groups, where dozens of properties of the models used have to be documented. In terms of the information model, key modifications from CIM1.5 include:

1. A model description now includes scientific domain descriptions, each of which can now be independently completed for a particular realm (e.g. ocean, atmosphere, etc.). Which means, for example, that the ocean realm experts only face describing the ocean properties and so on. It is no longer necessary for one person to marshal all the information.
2. Scientific domain descriptions now include all the key properties which are likely to be compared between model domains (resolution, grid extent, tuning properties, processes simulated, …).
3. The depth of information has been deliberately limited, with a focus on scientific descriptions rather than describing software details, both to simplify the documentation task in terms of scope, but also to force more concise descriptions of what is actually described and hence allow for more salient comparisons among CIM documents by the modeling community and by scientists using CMIP6 output.

The scientific domain descriptions are defined for CMIP6 via "specialisation" files which define the properties needed to have a configured model. They are organised following the eight official CMIP6 realms (as listed on the PCMDI github) – atmosphere, ocean, sea-ice, land surface, atmospheric chemistry, aerosols, ocean biogeochemistry, land ice – plus one top level specialisation. These specialisations constrain the basic CIM2 Model documents for these particular realms (by adding or removing properties as needed). These specialisations are defined and governed by ES-DOC realm experts who interact with the scientific community, and include several levels of reviews. This approach clearly separates the concepts encapsulated in the CIM from the specific activity it is used for (here CMIP6). Hence, using CIM2 for an another activity (e.g. CMIP7) will only require changing these specialisation files and not the underlying schema.

The specialisation files are written in python and form inputs into a set of generators that emit artefacts such as: mindmaps (for visualization purposes), JSON files (for configuration purposes), iPython notebooks or online questionnaire (for documentation creation purposes) ...etc. See section 5.1 below for more details on these tools. These files are available from https://es-doc.org/cmip6-specialisations.

Each realm process has an identifier associated with it which can be mapped to the CMIP5 model descriptions. This mapping allows for the auto-initialisation ("seeding")of CMIP6 model metadata from existing CMIP5 metadata holdings and thus significantly reduces the effort required from modelling groups.

Each group will fill up a model configuration file (CSV format, stored on github, created from the official CMIP6 source_id CV) which will describe for each realm the source: "blank" (start from scratch), "CMIP5" or other "source_id" (parent model). Initial model documentation will then be created by ES-DOC accordingly and handed to modelling groups for finalisation.

4.3 Ensemble and simulation

All simulations belong to an ensemble, even if there is only one ensemble member. Where there are multiple ensemble members, the ensemble description describes how the members vary along one or more ensemble axes (e.g. initialisation, forcing, …). The document creation workflow will ensure that the element identification code (variant id: r/i/p/f, see the

definitions in a related WIP document) in the data files is consistent with the enumeration along those ensemble axes.

In those cases where the same model is being used at multiple centres where they are each contributing to the same ensemble, one of the centres should complete an "uberensemble" record to explain how the sub-ensembles at the centres are linked.

The ensemble definition also provides the explicit links between the simulation and experimental requirements via conformances (see below).

Simulation and ensemble initial documents will be auto-generated from the contents of the ESGF archives with code running as part of the ESGF publication workflow (see section 5.1.3). Modelling groups will be required to edit these documents to fill in the information not captured in the global attributes, e.g. the Ensemble axis.

4.4 Conformance

The following a priori assumptions are made with respect to conformance information:

1. All simulation requirements are fully conformant
2. Conformance information is consistent with CMIP6 specifications

Therefore the modelling groups need only provide conformance information for those requirements to which they do *not* comply, to which they conform in a non-standard way or to which the governing MIP has made a particular request for information on how conformance was achieved.

ES-DOC will provide the tooling for modelling groups to provide non-conformance, non-standard conformance and required conformance information where required. The modelling groups should be able to provide this information relatively early i.e. before data submission. WGCM & WIP will establish a deadline for providing the information so that it can be used for the AR6 report.

4.5 Performance

The CMIP process is demanding of compute time, but it is not well known what resources are expended. Proper understanding of model performance coupled with the amount of effort needed to deliver the CMIP simulations will aid future planning. For that reason, it is proposed to collect information from the modelling groups and enter it into a "performance mip" using criteria established at a number of meetings and now published (http://www.geosci-model-dev-discuss.net/gmd-2016-197/).

The Performance CIM document is based on these criteria and may be attached to an Ensemble document - to give an indication of the average performance of all of its members - or to individual Ensemble Members if a single, ensemble-wide performance is not applicable to all members. Many ensembles will have the same performance for their members, so it is possible to attach the same performance description to multiple ensembles.

4.6 Other documents in Fig.2

- Responsible party: will be used to describe the group/people in charge of the experiments performed at the modelling institution.
- Citations: will be used to list any citations needed in support of documentation
- Machine: will list the types/models of super-computers used to run the models
- Online resource: is a URL provided and managed by the modelling groups for additional information (not captured in ES-DOC) or links to files, etc.
- Forcings appear in two different documents: in the top level specialization to capture the IPCC AR5 Table 12.1 properties and in the conformance document.
- Project: this is simply 'CMIP6'

4.7 CMIP6 controlled vocabularies

CMIP6 controlled vocabularies are being constructed by defining constraints on the various properties which CIM2 allows. CMIP6 core CV (such as the `activity_id` and `source_id` used to link documents) are obtained via the CMIP6 core CV repository[1] and associated documentation[2]. Other CVs are encapsulated in the CIM and the largest fraction is defined in the model specialisation documents described in section 4.2 (see example here).
The naming of the CIM documents is also based on the CMIP6 core CV.

4.8 Document life cycles and workflow

The lifecycle management of CIM documents is an important issue and requires specific software to support it. Once generated (either automatically or published by groups), it will be entirely managed by ES-DOC tooling. CMIP6 documents will be created, published and updated (including deprecated and replaced, with versioning, as documents are linked together after their initial production). An important innovation with CIM2 and the associated software upgrades is the idea that it is not necessary for the complete set of documents to exist before the first documents are published: examples will include the possibility to 1) publish the ocean description before any other domains, or even the top level model description, 2) publish the simulation descriptions during the ESGF workflow, without the accompanying further-info-URL additional target CIM documents (green boxes in Fig. 2) existing, etc. As shown in Figure 2, the linking will be made dynamically. The naming, IDs and versioning principles used to perform this linking are described below and illustrated in the workflows of the next section. [Note: Modeling groups may choose to deliver all documents pertaining to a single model in a single compound document, if desired].

5. ES-DOC software: for creating, linking, handling and using documentation

A variety of tools & web-services are needed to perform the actions in Figure 2.

5.1 Documentation Creation tools

To best adapt to the modelling groups internal workflows, several creation tools are proposed:

[1] https://github.com/WCRP-CMIP/CMIP6_CVs
[2] CMIP6 Key Controlled Vocabularies and Their Interdependencies (WIP white paper) https://docs.google.com/document/d/1N0pLdUA7_lgmK93MIQtdSeelHWPodJYOcWhDFDHiQ90

- **pyesdoc** – a python library supporting documentation creation, linking, validation, I/O, archival & publication. It also supports controlled-vocabulary management & simplified data request usage. It is extensively used within the ES-DOC eco-system. The UK Met Office plans to use this mechanism for metadata creation delivery.
- **questionnaire** – a web interface for direct user input of metadata required for the CIM documents; based on the CMIP5 questionnaire, but totally rewritten and rationalized for CMIP6 as well as much more modular. Can be initialised from the CMIP5 version of a model.
- **iPython notebooks** – creates a complete CIM document through user input to an iPython notebook. Can be initialised from the CMIP5 version of a model. GFDL and IPSL plan to use this mechanism.
- **Spreadsheets** - where appropriate spreadsheets can be used to gather documentation. The CMIP6 MIP's & Experiments have been documented in this fashion. Custom scripts, leveraging pyesdoc, convert the spreadsheets to valid CIM documents and publish them to the ES-DOC archive.

5.1.1 The ES-DOC Questionnaire

The ES-DOC Questionnaire is a web-based tool allowing users to directly enter content in order to create CIM Documents. It is based upon the CMIP5 Questionnaire, although the user interface is not as complex. A set of web forms is dynamically generated for a particular, potentially specialised, CIM document type. A user or set of users can then provide all of the information required by that document type. This is not expected to happen in a single session; Rather users will complete the Questionnaire iteratively. Once complete, the document can be published to the ES-DOC archive and viewed and/or compared along with any other CIM document.

Another use case for the Questionnaire is to allow users to "import" existing CIM documents into it and then alter information or add additional information manually before publishing the new document. There are two reasons for this functionality: 1) It provides a means for documents to be *mostly* auto-generated with small bits of human input added at the end (such as updates to the Ensemble document described in the section 4.3). 2) It provides a way for a similar document to be based on a copy of an existing document with just the different bits altered by a human.

5.1.2 Creation workflows: who is doing what when

Simulation descriptions will be auto-generated from the contents of the ESGF archives with code running on an ESGF data node as part of its publication workflow. Therefore no extra effort from the modelling institutes is required, above that already required to submit valid datasets to ESGF (**Fig. 4**).

On an ESGF data node, the raw information required to create the ES-DOC Ensemble, EnsembleMember and Simulation documents will be automatically extracted from the netCDF files and sent to the ES-DOC server. These raw descriptions will then be converted to CIM documents and published to the ES-DOC archive by code running automatically on the ES-DOC server.

The code running on the ES-DOC server will also initialize EnsembleAxis documents, which describe how the ensemble varies along the r/i/p/f axes of the ensemble. These documents will, however, need extra content added by the modelling groups to describe the nature of these variations, which are not recorded in the netCDF files (see Fig. 2).

Descriptions of models cannot be auto-generated, but they may be created by any of the tools, or a combination of tools, chosen by insitute preference, that are provided by ES-DOC (section 5.1).

The model documentation workflow (**Fig. 5**) starts at a modelling institute with the identification of an "ES-DOC officer", who acts as a liaison between ES-DOC and the modelling group. The ES-DOC officer is responsible for choosing appropriate document creation tools and then working with the home modelling experts who will record the actual model details. The ES-DOC officer can be a modelling expert but this is not a requirement, and it is likely that different people will be the experts for different model realms (atmosphere, ocean, etc.).

Once the documentation for a model realm has been completed, it is published to the ES-DOC archive by the ES-DOC officer. This process triggers a request back to the realm expert for a quality control review. If any corrections are required, then the document is republished to the ES-DOC archive and will be generally available at http://es-doc.org for viewing, editing via the Questionnaire and comparison with other documents.

It is possible that, within a modelling group, some realms may take longer than others to be fully documented. By splitting the publishing up into individual realms, parts of the model which have been completely described can be made available without having to wait for other realms which may not have been finished.

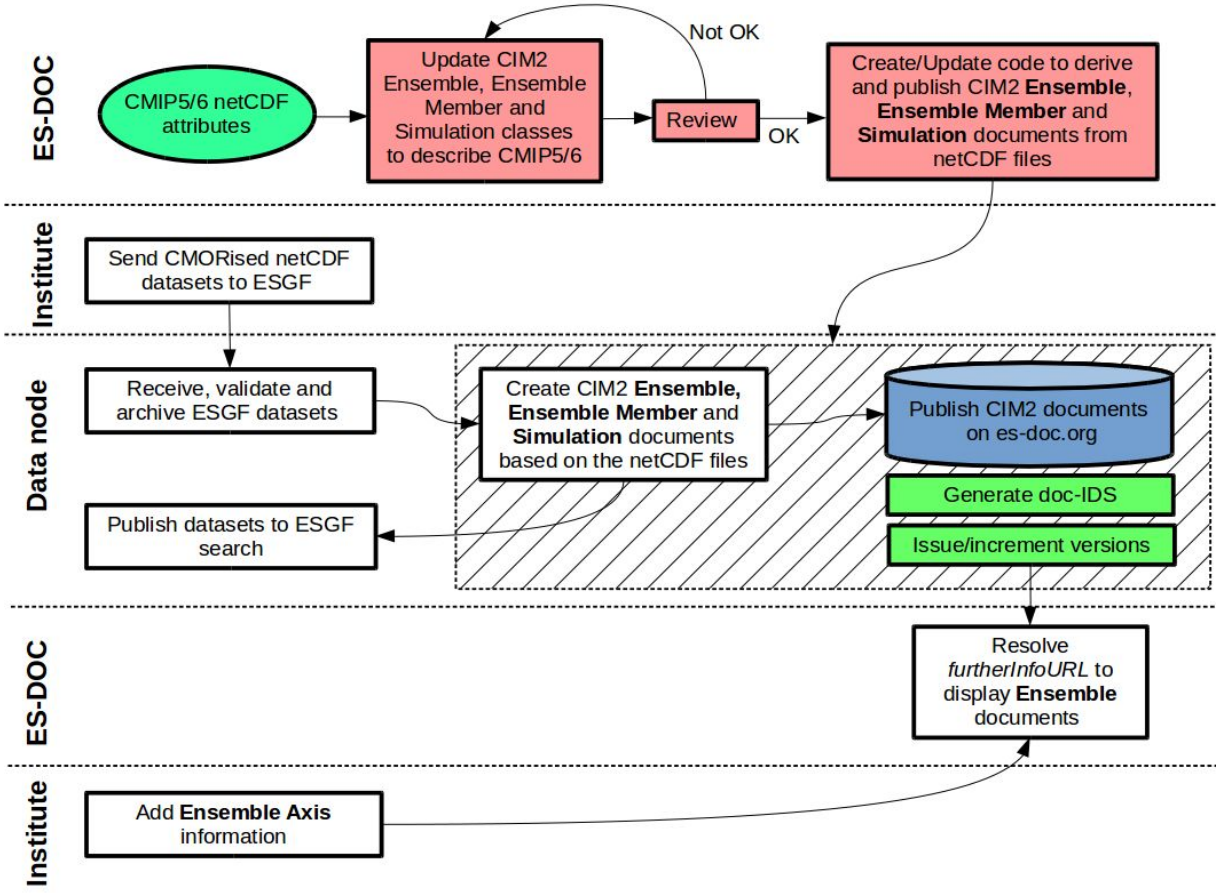

Figure 4. Simulation documentation workflow, distinguishing institute (i.e. modelling group) responsibilities from that of ESGF and ES-DOC. The top line of ES-DOC activities has been completed, and the other data node and ES-DOC processes will occur after netCDF files have been submitted to the archive.

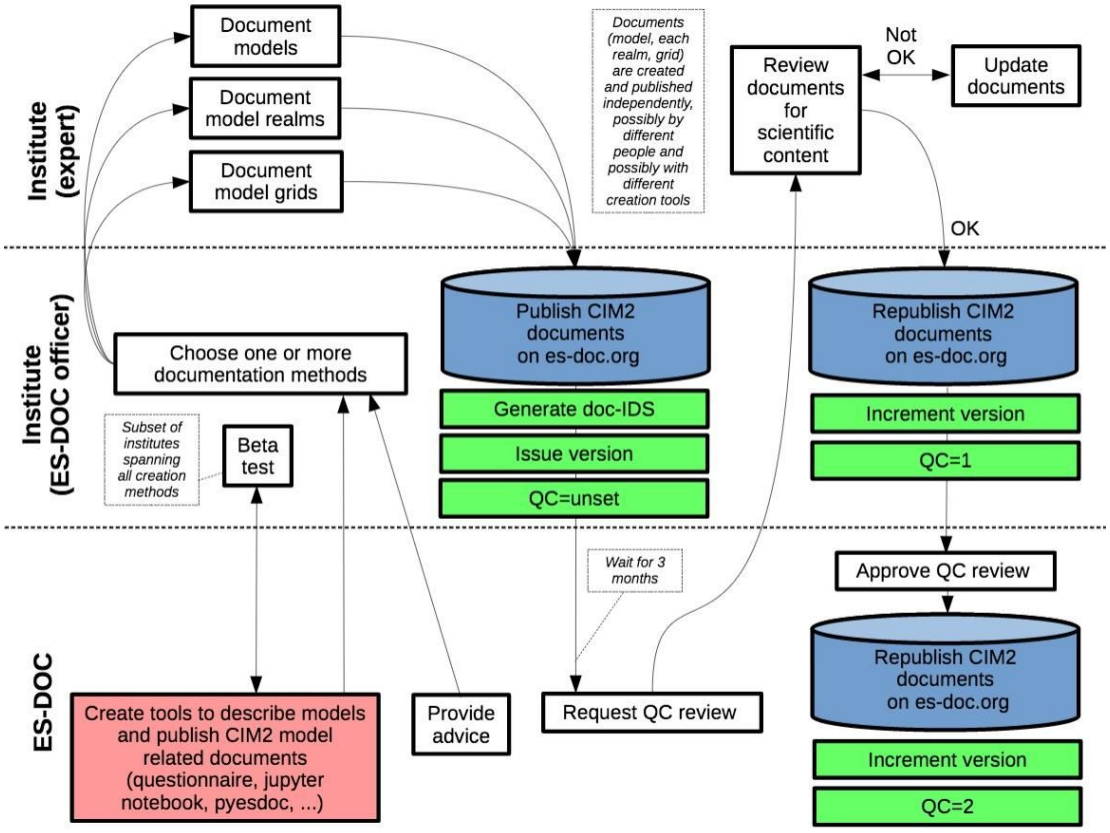

Figure 5. Model documentation workflow, outlining institute (i.e. modelling group) responsibilities and that of ES-DOC to jointly produce, publish and quality control the related CIM documents.

5.2 Linking it all together and the further_info_url

Locating a CIM2 document from a netCDF data file is done via the further_info_URL link which is part of the global attributes in the netCDF files. This is mostly automated or ES-DOC generated, as described below.

The further_info_url has the form:
http://furtherinfo.es-doc.org/<mip_era>.<institution_id>.<source_id>.<experiment_id>.<sub_experiment_id>.<variant_label>
(e.g., "http://furtherinfo.es-doc.org/cmip6.MOHC.HadCM3.historical.none.r3i1p1f1").

Further info URL's will land at a micro web-service maintained by the ES-DOC project. The web-service will validate the incoming further-info-url and redirect the user to a further-info-url landing page. The landing page will render the information harvested from the netCDF files created by the modeling groups and automatically recorded in the Ensemble CIM document. The page will present links to the full set of documentation available within the ES-DOC eco-system and potentially beyond. For example, the model name is stored as a string in the netCDF files, but will appear as a link in the further_info_URL page which takes a user to that model's CIM documentation. The documentation for each simulation of the ensemble, as well as the related citation (WIP White Paper: "CMIP6 Data Citation and Long-Term Archival", http://bit.ly/23rsG8l, Information: http://cmip6cite.wdc-climate.de) and

errata (http://wordpress.es-doc.org/web-service-dataset-errata/) services will also be accessed by links from this page.

An important feature of the Ensemble document is the ability to store the names of arbitrary web pages which are hosted elsewhere. This allows the modelling groups to link in repositories of information which are in entirely in their control, outside of the CMIP6/ESGF/ES-DOC ecosystem. The addresses of these web pages will appear in the further_info_URL page alongside all other ensemble information.

5.3 Model publication tables (aka "short tables")

A short subset of properties will be extracted from the CMIP6 specialisation files to form standard model description tables for modelling groups publications (as per WGCM request). The choice of these properties is defined with and review by the community in the same process as the definition of the specialisation files. The property subsets will be flat lists of specialization identifiers. Such identifiers will act as input to tools able to extract these subsets of properties and provide the output in differents formats (csv, pdf, …).

5.4 Use of documentation (view, compare)

The scientific community can search & browse all documentation published within the ES-DOC eco-system at http://search.es-doc.org. Model & Simulation inter-comparison will be supported via the ES-DOC comparator, see http://compare.es-doc.org.

6. Testing phases and interaction with the community

Testing is a key stage in the development of any software tool. Tests exists at two levels, a) unit tests which are part of the code base, which are run prior to a new release of the software, and b) functional beta testing, which are run after major features are added to the code. The functional beta testing focuses on the document workflow and therefore is broken down into generation, ingestion, viewing, and comparison steps. Since not all stages of a document workflow will be available at once, tests will be scheduled as development pieces are completed. The schedule and results of various tests are archived on github at: https://github.com/ES-DOC/esdoc-testing/blob/master/timeline_master_schedule.md

ES-DOC tools are supported via the es-doc-support@list.woc.noaa.gov mailing list. This list is monitored by key developers who are in a position to answer technical questions related to tool use. Scientific related questions are usually forwarded to applicable data or scientific realm experts.

Training on the various tools will primarily take the form of documentation on the ES-DOC wordpress site. Additional training may be available depending upon availability of resources at monthly coding sprints, through virtual webinars, or one-on-one tutorials.

All ES-DOC CMIP6 resources are available at http://es-doc.org/cmip6.

7. ES-DOC officer in modelling groups

The ES-DOC team invites modelling groups to appoint an ES-DOC officer to act

as interface between their institution and the ES-DOC team during the CMIP6 documentation process. This person will:

- be in charge of creating the institute-related CMIP6 documentation, which will involve interactions with scientists and IT experts.
- become familiar with various documentation methods that ES-DOC is providing for CMIP6 documentation
- advise on which documentation method(s) is/are best suited to the institution and provide internal training on how to use it/them
- manage any technical issues, liaising with the ES-DOC team as required

ES-DOC will in due course provide training and on-line support materials to ES-DOC officers so that they are able to carry out this role.

The ES-DOC person in charge of interacting with ES-DOC officers is David Hassell who will be the main liaison for this process during CMIP6.

8. The ES-DOC team

The ES-DOC team is composed of about 15 experts from multiple disciplines, with a long history of working together. Here is list of the current members, with their role and responsibilities:

- Chris Blanton (GFDL, USA): general beta testing
- Mark Elkington (MeOffice, UK): general beta testing
- Mark Greenslade (IPSL, FR): pyesdoc, Viewer and Comparator, base ES-DOC tooling, CIM2
- Eric Guilyardi (NCAS, UK and IPSL, FR): Coordination, realm expext (ocean, ocean biogeo.)
- David Hassell (NCAS, UK): Ensembles, Simulation, furtherinfoURL, linking it all together, link with ESGF (via Ag Stevens), official contact point for ES-DOC liaisons, realm expert
- Emma Hibling (MetOffice, UK): pyesdoc beta testing
- Bryan Lawrence (NCAS,UK): Coordination, CIM2
- Guillaume Levavasseur, Atef Bennaser (IPSL, FR): errata service
- Sylvia Murphy (NOAA/UCAR, USA, until Dec. 2016). Project management, testing coordination
- Charlotte Pascoe (CEDA, UK): Experiments, Conformance, realm expert (atmos, atmos chemistry,..)
- Ruth Petrie (CEDA, UK): realm expert (Sea-ice)
- Martina Stockhause & Hans Ramthun DKRZ, Germany): data citation service, link with further-info-URL
- Allyn Treshansky (NOAA/UCAR, USA): Questionnaire, CIM2, project management support

The ES-DOC team has a telco every week on Monday and its developements and telco minutes can be found on https://github.com/ES-DOC and on COG: https://www.earthsystemcog.org/projects/es-doc-models/

The ES-DOC PIs regularly review the work of the ES-DOC team, provide guidance and are in charge of identifying resources for the team. Current members are:

- Balaji (GFDL, USA), WIP link
- Cecelia DeLuca (NOAA/UCAR, USA)
- Sébastien Denvil (IPSL, FR)
- Eric Guilyardi (NCAS, UK and IPSL, FR)
- Bryan Lawrence (NCAS, UK)
- Karl Taylor (PCMDI, USA), WIP link

9. Development status and timeline

The current target for public release of the CMIP6 documentation tools and services are early June 2017.

Appendix I: CIM design and governance

Capturing the concepts in the CIM

In CMIP5 version 1.5, CIM documents were selected from one of the "activity" (**Experiment**, **Simulation**), "software" (**Model**), "data" (**Data**), "shared" (**Platform**) and "grid" (**Grid**) packages (where the boldface denoted particular document types). Apart from the experiment definitions which were constructed directly in XML, all the other documents were created via the online CMIP5 questionnaire (a complex Django application, now retired). The sheer volume of information required to fill in the CMIP5 questionnaire was intimidating, and the necessity to have it all before publication was inhibiting, hence the design criteria for CMIP6 to make the infrastructure less intimidating, and to streamline production by reducing complexity and duplication and increasing automation.

The revised document structure in CIM2 (Figure A1) has entities at a finer granularity, some of which can be generated automatically, and all of which are themselves much simpler. In addition, some of the entities are exposed so that they can be created once, and simply pointed to from other documents, rather than re-described many times. By design there is no necessity for documents to be created in any particular order, and the connections are made by controlling the terms used to name entities. The canonical representation of the CIM 2.0 schema is a simple pythonic format, in contrast to the more complicated CIM1.x approach where the canonical representation was UML which was then transformed into convoluted XSD, along with a custom mindmap format existing alongside the UML. This CIM2 simplified pythonic format simplifies downstream tooling chains while keeping al the concepts of CIM1.x.

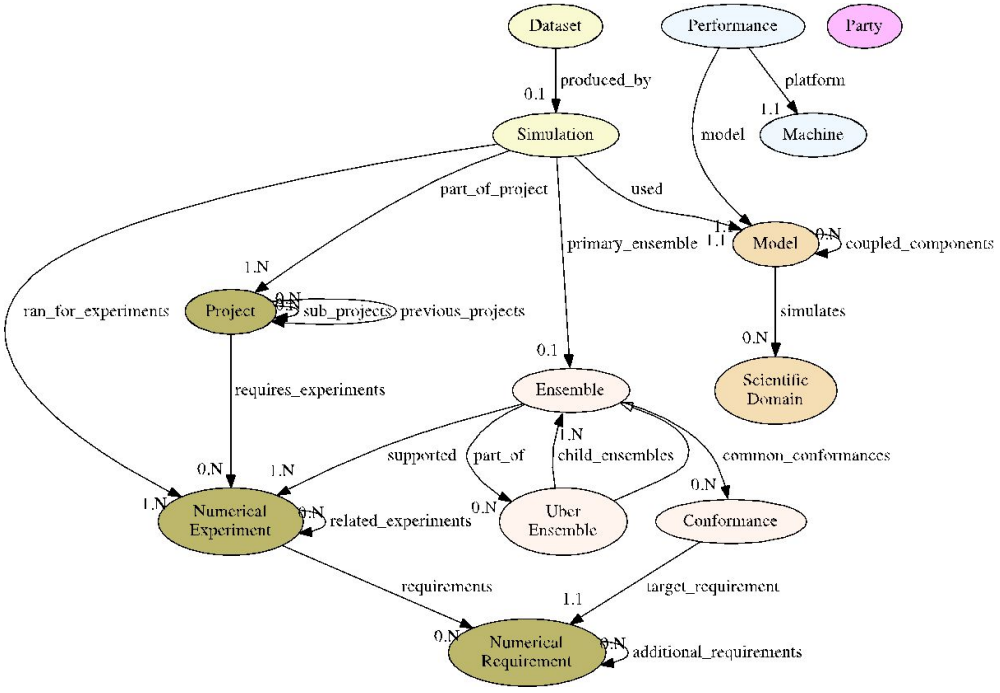

Figure A1: Set of key CIM2 documents of relevance to CMIP6 and their relationships. Different colours denote different packages: from the khaki of designing the experiments, to the yellow of the datasets and their auto-generated simulation descriptions, the beige set of ensemble and conformance documents which will need to be created to make the links between ensemble

members and experimental requirements, the orange model descriptions, and the blue descriptions of model performance. Following the ISO nomenclature, individuals and organisations are described using party descriptions.

The CIM is defined using a simple python formalism which exposes the key aspects of the Unified Modelling Language (UML) via a prescribed dictionary format. All of the ES-DOC machinery is now built on top of these definitions.

Short overview of the key CIM2.0 packages:

- **Designing**: includes all the classes associated with describing projects, experiments, and their requirements (section 3.2). It will be seen that there are now many more documents associated with designing experiments, but this proliferation aids re-usable requirement definition, and will in any case be handled before CMIP6 begins.
- **Data**: now includes both the dataset and primary simulation documents, both of which used to be manually constructed, but will now be automatically generated during ESGF publication. Most of the information about why simulations exist was already implicit in the DRS and file headers[3], and the remainder has either been added into the file requirements, or split out into a complete changed activity package.
- **Activity**: Now consists primarily of the information that distinguishes simulations apart within an ensemble, and how they conform to experiment requirements (for example, which datasets were used to meet forcing requirements). Because many of the conformances are likely to be either re-used between experiments, or will be associated with variations along ensemble axes they can be entered either in re-usable standalone conformance documents, or as part of ensemble descriptions.
- **Model** (class in **Science** package): In CIM2.0 (unlike with previous versions) the concepts of scientific and software descriptions have been clearly separated. For CMIP6, we only worry about components where a model consists of fully coupled standalone components, but the primary description is via standalone descriptions of the scientific domains simulated by the models, rather than their underlying software confiurations.
- **Performance**: New to CIM2 is the performance package, which includes the platform description from CMIP5, made more useful by improving the information required along with the performance criteria requested for the Performance MIP [ref]
- **Party** (class in **Shared** package): in CIM2, individuals and organisations are described using the ISO standards formulation for a party, linked to ORCID identifiers.

The CIM1 Grid document has been deprecated (because of its complexity) and what was previously described by it is now split between the Data and Model packages.

The current CIM2 repository can be found on GitHub.

CIM governance principle can be found here:
https://github.com/ES-DOC/esdoc-cim/wiki/CIMGovernance

3 https://docs.google.com/document/d/1h0r8RZr_f38egBMMh7aqLwy3snpD6_MrDz1q8n5XUk

References:

- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, doi:10.5194/gmd-9-1937-2016, 2016.
- Guilyardi E., B. Lawrence,V. Balaji, S. Callaghan, C. DeLuca, S. Denvil, M. Lautenschlager, M. Morgan, S. Murphy, K.Taylor and the Metafor team (2013). Documenting climate models and their simulations. Bull. Amer. Met. Soc., 94, 623–627, doi: 10.1175/BAMS-D-11-00035.1
- Lawrence, B. N., Balaji, V., Bentley, P., Callaghan, S., DeLuca, C., Denvil, S., Devine, G., Elkington, M., Ford, R. W., Guilyardi, E., Lautenschlager, M., Morgan, M., Moine, M.-P., Murphy, S., Pascoe, C., Ramthun, H., Slavin, P., Steenman-Clark, L., Toussaint, F., Treshansky, A., and Valcke, S. (2012): Describing Earth System Simulations with the Metafor CIM, Geosci. Model Dev. Discuss., 5, 1669-1689, doi:10.5194/gmdd-5-1669-2012.