CMIP6 Quality Assurance

Authors: Frank Toussaint, Martina Stockhause, Michael Lautenschlager

v1.0, Date: 2015-12-17

v1.1, Date: 2015-12-18 integration of comments of KT

V1.2, Date: 2016-03-07 integration of comments from MJ

V1.3, Date: 2016-04-03 integration of comments from SD and others

Scope

This document proposes and discusses different aspects of data and metadata quality and quality assurance during the CMIP workflow.

It covers the data production and storage path from the data producer to the final long term storage of stable data in the IPCC DDC, including technical needs within the ESGF data infrastructure as well as data quality policies. It does not cover quality questions of adjacent metadata as metadata on models, errata, experiments, simulations, citation information etc.

Content

1. **Executive Summary**

For CMIP6, about 50 times more data than for CMIP5 are expected. With the background of the CMIP5 experiences various aspects of data and metadata (MD) quality should be taken into account:

-- ***Clear file metadata requirements for publication***

This document sets out clear file metadata **Data Publication Requirements** such that:
(1) Compliance with mandatory requirements from controlled vocabularies, data reference syntax and NetCDF/CF global attributes (publication blockers) can be easily checked by the data providers (M1/D1 in Figure 1);
(2) Data with publication blockers will not be published (M2/D2 in Figure 1 below).
The complete quality assurance process will go much after ESGF data publication.

- ***Documentation of Quality: Comments and Results of Checks (Impact on Modelling Groups & CMIP6 data management)***

The community should agree on checks as given in the figure in section 2. For all checks on data and metadata and on their accessibility inside and outside the CMIP project there should be agreement on high transparency regarding the checked criteria and their results, e.g. by xml files distributed by an atom feed.

- ***Consistency of Data (Impact on ESGF & CMIP6 data management)***

Within CMIP5 the identification of data objects by the DRS names was not reliable because of different vocabularies used by the different infrastructure components and inconsistent data versioning at some data nodes, which caused data inconsistencies within the ESGF between data and their replica. Data consistency for CMIP6 is required for data services across data nodes, and so the DRS must be backed by reliable controlled vocabularies of system wide accessibility.
As an additional step to data homogeneity we recommend to use a common data output software and ensure community agreed quality requirements. In this sense we encourage the use of the CMOR software.

- ***Application of Software (Impact on CDNOT & CMIP6 data management)***

The quality of the ESGF process and the user guidance its tools can be improved. It presently allows for different ways to publish and unpublish data, depending on parameters. Here the first inhomogeneities between the data nodes occur. For this and other cases automated continuous consistency checks on data nodes of data and the describing MD would be very helpful. Furthermore, it needs a clear versioning. CDNOT should make sure that data nodes install new versions within a reasonable amount of time. If some partners do not have the resources to install and maintain the programmes, we propose cooperation, i.e. nodes run by one or more partners. A classification of nodes by up-times and technical performance might help, too. Some feedback from colleagues suggest, that training courses on usage of ESGF can improve conformance.

For all data infrastructure it is required that it has reached the productive state and the components are no longer under development and testing. Workshops and/or tutorials for the uniform use of the system are necessary with respect to ESGF data publication as well as ESGF data search and access

- ***Data Management Planning (Impact on CMIP6 data management)***
  As measures for data consistency and homogeneity need to be observed by all data producers, one should agree on a common data publication policy of the data nodes for at least egregious cases of violation of the Data Management Planning's rules. Some essential parts of the centres' checks (e.g. DRS) should be integrated into the ESGF publication process and abort it in case of deviations (see M2/D2 in chapter 3.2 and 3.3.2). This can stabilize the ESGF publication of homogeneous data. On the other hand, a standalone tool for M2 checks can be corrupted by improper handling or code change.

At the end of every chapter in the following text, the most important suggestions are listed. All of them are listed in the Table 1 below organized under subheadings according to who has primary responsibility.

Table 1: summary of requirements and recommendations, proposed by the authors.

Complete List of Proposals by the authors Recommendation/Requirement (for a table of requirements to files see 3.3.2)	Section
Data Centres (ESGF Data Nodes and Long Term Archives)	
archive the results of checks of data and make them available together with the data	3.2
check the MD and the citation information, a final check by the authors is recommended.	3.3.4
ESGF	
Perform extensive debugging checks and make available debugging tools before each software release; appoint someone specifically responsible for ensuring bugs found after release are promptly corrected. A test suite like the one proposed by IPSL for automated ESGF debugging would be helpful (see below).	3.4
Enable community annotation of the data	3.2
provide the files with persistent identifier syntax as tracking ID,provide the files with persistent identifier syntax as tracking ID	3.3.1

Do the essential parts of the checks at ESGF publication (M2) at the ESGF publication process and inhibit publication in case of severe violations to the agreed rules	3.3.2
do checks of the data values, if any, in a transparent way, well agreed by the user community at publication	3.2
use a sensible transparent arrangement of the elements of the data checks and make the output easy to read for third parties (i.e., users of the data catalogue)	3.3.2
review and stabilize relevant softwares’ release processes and transparently prioritize the development of the different tools. This may include a central point where enhancements can be proposed	3.4
establish automated continuous consistency checks at data nodes on the data and the MD. The necessary corrections need to be carried out	3.5
CDNOT	
make sure that CMIP relevant data nodes install new software versions within an agreed, reasonable time, e.g. a month	3.4
make clear, that at the time of ESGF data publication the contents and structures of file names and global attributes are as agreed[1] by whatever software the data have been generated	3.3.1
harmonize the metadata checks done by the different data nodes	3.3.2
commit to keeping DRS conventions stable (or at least backwards compatible) during the project. This includes CVs and the handling of versioning. Except in extreme cases (to be agreed by CDNOT and the WIP panel) a change of these conventions should be avoided,	3.1
agree on a procedure to accelerate the throughput through bottlenecks, i.e. define a prioritised set of data/checks or to decide, that the data producers might conduct the checks by themselves. To define here a subset of data is preferable, as it not only yields for bottlenecks at the time of checks but also for shortages at other parts of the workflow in CMIP6	3.1
point out clearly for all specifications whether they are mandatory, recommended or desirable	2.2

agree on a fixed scheme of identifiers which is mandatory for MD and adjacent[1] MD equally	2.2
agree on common measures to be taken when a data node violates important parts of the publication policy	2.2
define in detail the application steps and parameters of the ESGF software	3.5
decide, e.g. on basis of a statistics of usability and downtimes of the data nodes, on support offered to them	3.5
offer workshops and/or tutorials to administrator users of the ESGF software	3.5
not have the M2 metadata checks done by the data producer only. If they are conducted there, they should be repeated at the repository's site, if possible,	3.3.2
Data producer	
use person IDs like ORCID for references to persons (authors, editors, investigators…) in addition to the person names, wherever possible.	3.1
Emphasize the importance of updating the tracking ID. Stored as a global attribute in a file whenever anything is modified in the file.	?
Miscellanea	
develop and publish quality assurance criteria and their results for adjacent MD (ES-DOC?)	3.
make the status (in work, ready…) of an adjacent[1] MD object's instance available via API for a merge of the data	3.
make the CV conveniently available via internet in different formats to facilitate software use of the authoritative lists (data centres?)	3.1

[1] Adjacent I will call MD that don't describe the data object itself like added errata or comments, MD about the model or the simulation, citation information etc. They may be on different granularity layers. So care has to be taken to enable their technical linkage to the data MD later.

2. Introduction

2.1 Objective

For CMIP data, distributed production, multiple replication, and dissemination by different data nodes lead to an interdisciplinary scientific and engineering system that calls for special means of steering and control. An industry approach to solve these challenges is Systems Engineering which, however, cannot be directly transferred to globally distributed research architectures but many elements of this field of engineering may be applied to CMIP6.

In addition, the expected high volumes of data produced and administered in this project force a high degree of automation in data processing, cataloguing, and dissemination. This in turn requires data homogeneity which in federated projects is difficult to achieve.

This paper discusses some aspects of data and metadata quality and quality assurance during the CMIP workflow. It has the intention to have solutions for some of these points implemented into the CMIP6 process which here is structured into the following steps (see also figure below).

- *Data production and post-processing phase:* The data still is in the custody of the data producer who is responsible for quality issues. Data objects are finalized by a software package which is common at all contributing sites (CMOR).

- *Publication phase:* The data are in the hands of the ESGF data node operators for publication (some nodes may be run by data producers);

- *Project Phase:* First dataset versions are available in the ESGF for download. Data is shown in presentations to a selected audience or within the CMIP6 project.

- *Community Phase:* As the scientific community starts to analyse the data, individual datasets might get revised and published as new versions in the ESGF. For CMIP5, papers were submitted and published within this phase, and the IPCC AR5 was written without data citations. For CMIP6 data citations should be integrated in these papers as verifiable collections of certain versions of datasets (early data citation reference). NB: To some degree at least, the Project Phase and Community Phase overlap.

- *Bibliometric Phase:* Towards the end of the project, data becomes more stable and is placed in a long-term archive (LTA) for interdisciplinary reuse, e.g. by the IPCC DDC AR6 users (LTA reference data come with DataCite data publication and DOI minted).

2.2 Data Management Aspects

A Data Management Plan is the project's internal agreement on use of external and internal standards and on policies.

For projects of many partners, a Data Management Plan has benefits. In the partners' view the obligations incurred as part of the agreement will be more demanding. In case a strong necessity of any changes of data structure or formats arises, it clearly demands a new agreement on these standards. So it helps to prevent that community from unilateral changes of the agreed specifications.

In CMIP5 various data management standards were specified but not always followed. A coherent rejection policy of the data nodes will be crucial. At DKRZ the experiences with CORDEX data are promising. In general data producers were not hesitant to do the corrections in question. In CMIP, however, the repositories are more distant from the data producing institutes.

In CMIP6 different aspects of data management are discussed in white papers which might lead to distributed definitions and project standards. So for CMIP DMP might refer to Data Management Planning rather than to a single, monolithic Data Management Plan. This includes the different White Papers where the project's practices are laid down.

Some aspects of a typical planning of data management and related (White) Papers

Data and file format	NetCDF-CF Standard
File names and global attributes	*CMIP6 File Names and Global Attributes*
DRS	*CMIP6 Data Reference Vocabularies*
Variables	coordination with MIPs
Archive contents	*CMIP6 Data Citation and Long Term Archival*
Access Policies	*CMIP Licensing and Access Control*
ESGF data publication policies	*this WP and WIP / CDNOT*
Data Replication	*CMIP6 Replication and Versioning*
Data access/data sharing/ToU	*CMIP6 Licensing and Access Control*
Data Citation units and policies	*CMIP6 Data Citation and Long Term Archival*
Data quality assurance	this Position Paper
Possible remarks by data creators	*CMIP6 Errata for CMIP6*
Versioning	*CMIP6 Replication and Versioning*

It is proposed for the planning of data management to…

- point out clearly for all specifications whether they are mandatory, recommended or desirable,
- agree on a fixed scheme of identifiers which is mandatory for metadata (MD) and adjacent MD equally,

- agree on a common measures to be taken when a data node violates mandatory parts of the publication policy

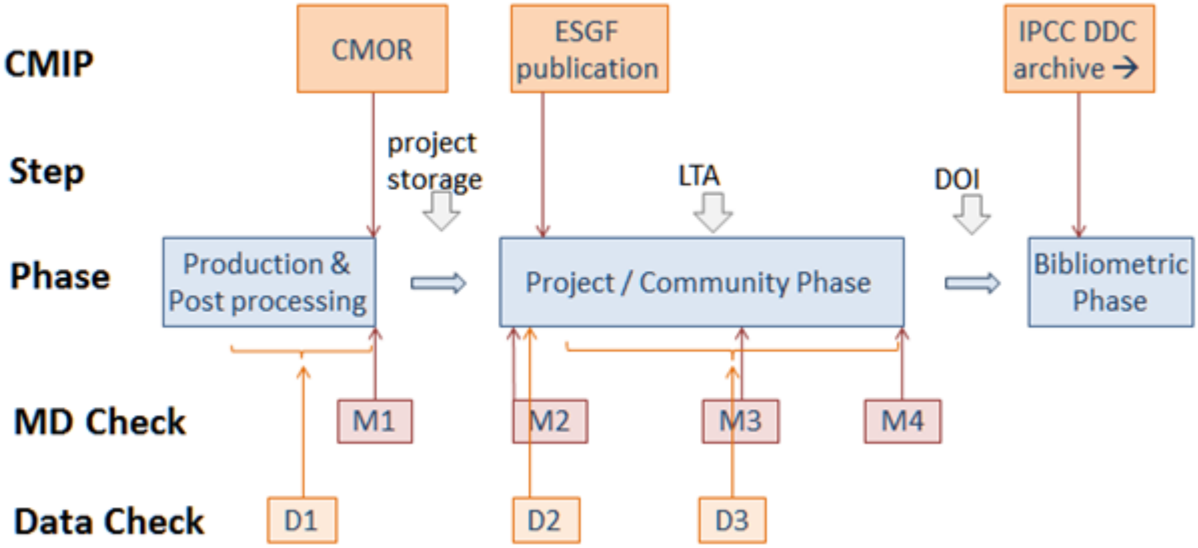

Figure 1: Rough workflow for a data package in CMIP

(D1,D2.. Data Checks, M1,M2.. MD Checks, QC of software not represented)

3. Data, Metadata, and Software Quality

With respect to necessities and responsibilities of the institutes and data centres, Quality Assurance (QA) for data differs from QA for metadata (MD). For the data values its creator (author/editor) has the main responsibility for correctness and scientific quality. Here the concept of quality depends on the use of the data – so it has to be related to the accompanying MD which contain information on what the data is adequate for. As MD is concerned, the data centres partly have the responsibility for QA. They should keep track of, e.g., completeness and comprehensibility. Within a project the agreement on these responsibilities should be part of the data management planning.

In this chapter data, metadata and software quality aspects will be discussed. Quality checks of adjacent[1] metadata like CIM[2] model or simulation descriptions, citation information and errata need to be integrated into these processes. Though the specifications of adjacent MD is not part of this chapter we propose that for them at least some quality aspects should be fixed.

It is proposed for the planning of data management to…

[2] The Common Information Model aims to describe, e.g., models and simulations by selected MD.

- develop and publish quality assurance criteria and record their results for adjacent MD,
- make the status (in work, ready…) of an adjacent MD object available via API to relate it to the core MD. This has implications to the producers of these MD like ES-DOC, users of QC tools etc.

3.1 Reliable Identifications and Linkages: CV, PID & More

In the planning of data management a couple of aspects can be considered that make quality assurance easier. For big amounts of data automated quality control is essential. For example, header entries (in NetCDF mostly key/value pairs) can have checkable or uncheckable contents: A typo in a string containing a contact person's name will be in the file headers probably longer persistent than the data are, as fields with foreign names or comments often stay unchecked. For data in free text fields (e.g. names), a list of controlled vocabularies (CV) can not solve this either. Here some persistent identifiers (PID) on persons like ORCID allow for identifying wrong strings by their check digit. This is especially important for identifiers that are compounds of abbreviations or other meaningful parts like the DRS syntax.

Controlled Vocabularies (CVs) are central standards of the project. To enable their use by all partners and their software tools, low-threshold access is essential, e.g. by an http accessible repository. Avoiding unnecessary definitions can make handling and discussions on CV elements easy: no case sensitivity in CVs.

For projects with high data volumes like CMIP6, a bottleneck might be introduced into the job stream if every one of the metadata checks had to be completed prior to publication of the data. Nevertheless all mandatory **Data Publication Requirements** (publication blockers) must be met before publication. This is especially valid for M2 and combined D2/M2 checks.

There are at least three possible solutions to bottlenecks at M2 checks:

1. reducing the amount of data:
it might be advisable to define a data core that is prioritized in all data checks,

2. reducing the amount of checks:
it might be advisable to define a core of items to check that is prioritized in all data checks,

3. distribute the checks:
one might prepone the data checks (M2) to the data producers and rely on their results.

As a procedure to accelerate the throughput through bottlenecks, we propose to consider reducing the number of MD checks performed to a smaller core set if performance problems are encountered (2. above).

It is proposed for the planning of data management to…

- commit to keeping DRS conventions stable (or at least backwards compatible) during the project. This includes CVs and the handling of versioning. Except in extreme cases (to be agreed by CDNOT and the WIP panel) a change of these conventions should be avoided,
- urge data producers to use person IDs like ORCID for references to persons (authors, editors, investigators…),
- make the CVs conveniently available via internet in different formats to make sure, software refers to the authoritative lists wherever possible,
- agree on a procedure to accelerate the throughput through bottlenecks, i.e. define a prioritised set of data/checks or to decide, that the data producers might conduct the checks by themselves. To define here a subset of data is preferable to defining at subset of checks, as it not only yields for bottlenecks at the time of checks but also for shortages at other parts of the workflow in CMIP6.

3.2 Quality of Data Values

The first group of checks of the data values (see figure chapter 2: D1) is usually done during the data production process. This allows finding principle errors in the data generation process, so this part of data quality assurance is mainly up to the data creator.

Depending on the project's internal agreements (planning of data management), a second check on the data contents (D2) at the distributing data centre (DC) may be desired. Here the files often undergo more general checks of their structure and MD contents, too. Hence they are usually opened and additional checks can be easily added. In general, here the data repository can offer a supporting service. In CMIP these data checks are partly integrated into the checking tools of, e.g., DKRZ and BADC. This information should be archived in a normalized and easy accessible form together with the long term archiving of the data.

The proof of the data is the publishing[2], so the final step of data assessment (D3) occurs in the scientific community – in the project and in the public. They should be annotated and accessible to the data users at data access, as this information influences the data's user dependent fitness for use. Here too, the data centres can support by allowing for comments on the data.

Definitions: who - where - what:

D1: Data creator - creator's site - checks during and after production

D2: DC in cooperation with data creator - data centre's site - e.g., outliers, validation against thresholds …

D3: community - e.g. data centre - annotations on all relevant topics including errata

It is proposed for the CMIP planning of data management to…

- enable the M2 MD checking tools to include D2 data checks on request in an easy modular way and offer this as a service to the data producers. This, however, will affect the performance of M2/D2 checks.
- do checks of the data values, if any, in a transparent way, well agreed by the user community,
- archive the results of possible checks of data and make them available together with the data,
- allow for annotations on errata to the data.

3.3 Quality of Metadata

Various kinds of metadata will in CMIP6 be included within the same files as the data and this will be augmented by metadata hosted elsewhere. Metadata are important for the informed reuse of the data. In a distributed environment metadata (MD) can be distributed too, and need common identifiers to give a merged view in the presentation to the user. In case of CMIP metadata we have used MD from the file headers and data nodes, collected CIM MD, citation information, errata and annotations as well as the results of quality checks – most of them at different locations. These data sources often produce their output on different granularity levels, i.e. on different levels of the DRS tree. Here only a sophisticated workflow can ensure correct assignments between the various data objects. In CMIP, however, there is not yet a common workflow in place. The project lacks interfaces between some of the different metadata sources. The ESGF Working Team on Data Quality presently tries to find a common workflow to integrate this meta information.

3.3.1 M1 – Checks & Homogeneous Metadata Output at the Data Producer's Site

Some of the different intentions for the use of MD are

- for the data creator: to get credit for the data by providing citation information, and to support efficient transport to the data centres, and to the end users.
- additionally for the data centre: to make the data homogeneous and searchable,
- additionally for the data user: to assess data with respect to usability/fitness for purpose.

So the first MD checks (M1) should at least support stable replication from the producer to the data centre. From the ESGF data publication on, the data objects are expected to be stable which may be controlled by, e.g., checksums. Corrections only should occur in connection with a new data version. So M1 also needs to cover all file header information as agreed on during the planning of data management, which at least comprises the use MD[3]. A common output software as e.g., CMOR can support the data homogeneity at this stage. In previous projects it turned out that for some institutes, e.g., long author lists were not stable during the project runtime. This was, e.g. due to changes of work contracts or, happy enough, due to a marriage. The appropriate use of CMOR, i.e. change of file header fields as necessary during the project run, will put the responsibility to the data producer.

At the data centre the MD mostly is used for cataloguing, so general topics for data search (facets) need to be stable, too. In addition, an project internally stable identifier should be added reliably. Here the simple file name does not yield, as during transport processes touching of file names cannot completely be excluded – which is difficult to check.

It is proposed for the CMIP planning of data management to…

- provide the files with persistent identifier syntax as tracking ID,
- make clear, that at the time of ESGF data publication the contents and structures of file names and global attributes are as agreed[4] as well as the PID string in the file header[5] by whatever software the data have been generated.

There will be a table in the WIP White Paper *CMIP6: File Names and netCDF Global Attributes* listing the checks to be performed. These checks can in addition already be conducted at the creator's site during data production to avoid unnecessary data reprocessing.

3.3.2 M2 – Checks of the Metadata Before ESGF Publication

Items listed as "Required" in table 2 must be checked and verified as correct before ESGF publications. It will considerably enhance the value of a provider's contribution if they are also able to comply with the "Recommended" items before publication, but we understand that this may not always be possible. The M2 checks will include additional checks beyond the "Required" and "Recommended" checks, designed to give a fuller analysis of the metadata quality. The full extent of these checks may evolve in response to experience and feedback.

The results of M2 are of interest for data users as well as for technicians who need to manipulate them via scripts. For both a clear arrangement of these data is essential. This includes in-file explanations of abbreviations, parsable text structure and MD on these quality data as, e.g., timestamp and person responsible for the checks. References by URL are not advisable, as their stability is doubtful.

At M2 there is mutual interference between the need for a detailed check and quick publication. So a project might agree on doing M2 or parts of it at the data producing site. Unless remote checking is used, even a partial merge of M2 with M1 into one software package, however, breaks the rule that checking should be conducted by another party. The Data Publication Requirements (see table 2) (review of DRS+filename) should be integrated into the ESGF publication process and block it if the checks of requirements are not passed. This avoids technical interrupts at later stages of the CMIP6 data workflow.

At WDCC/DKRZ a new version of a checking tool for CMIP data has been developed within the EU Project IS-ENES-2. Previous versions have been used for data of projects like CMIP5 and CORDEX. The latest version is available at https://github.com/IS-ENES-Data/QA-DKRZ Documentation at http://qa-dkrz.readthedocs.org/en/latest/.

The CEDA Compliance Checker (CEDA-CC http://proj.badc.rl.ac.uk/exarch/wiki/PackageFCC) has been used for CORDEX, and is in use for SPECS and CCMI and the ESA Climate Change Initiative. CEDA-CC has been developed to ensure that it can easily be run by data providers to check all files before sending data to archives for publication.

An example for M2 checks presently conducted at WDCC is the checklist of the CORDEX project: https://github.com/IS-ENES-Data/esgf-cordex/blob/master/CORDEX_qc.xlsx . For the checks recommended for CMIP6 see the table above, mostly taken from WIP White Paper *CMIP6: File Names and NetCDF Global Attributes* and the accompanying documents.

It is proposed for the CMIP planning of data management to…

- use a sensible transparent arrangement of the elements of the data checks and make the output easy to read for third parties (i.e., users of the data catalogue),
- not have the M2 metadata checks done by the data producer only. If they are conducted there, they should be repeated at the repository's site,
- harmonize the metadata checks done by the different data nodes,
- integrate the essential parts of the checks at ESGF publication (M2) into the ESGF publication process and inhibit publication in case of severe violations to the agreed rules.

3.3.3 M3 – MD Checks as Technical Quality Assurance (TQA)

After the project's data production phase, long term archiving (LTA) of the reference data starts. This has the implications of long term availability of and open access to data and metadata.

Now the data is not only distributed within the project but to a wider community. The checks at this point refer to such things as data accessibility via the means given in the MD, links/access to provenience MD and other meta information like annotations. References to other data like describing journal texts might be checked, too.

In this phase the consistency of the MD information can be enhanced by cross checks against adjacent MD objects like CIM documents or other external information (Technical Quality Assurance, TQA). However, in CMIP unlike in some other projects, project phase and community phase of data distribution are not well separated. The data are accessible right after ESGF publication – inside and outside the project.

For CMIP5 a Technical Quality Control process for metadata and data was developed which is described at http://redmine.dkrz.de/collaboration/projects/cmip5-qc/wiki/Qc_l3#Criteria- for-QC-L3DOI-publication . The passing of some of the checks is required, whereas the results of others are logged or documented. The checks include double-checks of the data and cross-checks between data and metadata. In case inconsistencies are found, the data author is contacted for clarification. No scientific judgement will be passed on CMIP data during the QC process. Such an assessment is clearly the responsibility of the modelling groups.

It is proposed for the CMIP planning of data management to…

- check at start of LTA all aspects of data consistency, e.g. completeness, accessibility, checksum/size, match to MD etc.

3.3.4 M4 – Author Approval and Final Checks

Finally the data becomes used outside the project by scientists of other fields and by the public. As in CMIP5, we for CMIP6 also propose that at the end of the processing chain, the author be given the opportunity to finally review, enhance, and complement the metadata within a reasonable period of time. In this step of authors' contact, the focus is on, e.g., exact authors' list, possible scientific quality assurance (SQA) of data done by the producer, and possible amendments of references to publications related to the data.

After final approval of the citation MD, these are fixed (see the White Paper *CMIP6 Data Citation and Long Term Archival*) and the DOI is registered at DataCite together with the citation MD.

It is proposed for the CMIP planning of data management to…

- check the MD and the citation information, a final check by the authors is recommended.

3.4 Quality of Software

The quality of software used in a project should mostly be transparent to the end-user. However, it strongly influences the quality and so the usability of the data. Some companies have standard systems like the Capability Maturity Model (CMM) to optimize their software generation processes. Others rely on good practice.

In the CMIP process, there are not yet detailed internal rules for software quality. This applies not only to ESGF as the main part but also for the many other CMIP tools in connection with CIM/ES-DOC, quality checks and CMOR. Some programmes are regarded as prototypical by their users whereas the software specialists already concentrate on other work – which is certainly not less important. "Broken by design" are the words of a scientist when he heard that in ESGF a deletion of files is possible which is unobserved by the system. In addition, a clear software versioning marked in the output is needed.

The ESGF data dissemination system is highly fault-tolerant with respect to the data to release. This normally would be a big advantage for data publishing and data propagation as long as there was no need for homogeneity. In CMIP5, however, this convenience turned into its opposite. The bulk data's diversity in the metadata made it difficult to handle them by machines, i.e. to write tools to do so.

Another reason for data inhomogeneity was the use of different software versions. It should be made sure that the data nodes install new ESGF software versions within an agreed, reasonable time, e.g. two weeks. Actually, in CMIP5 there were data nodes that delayed the installation of new software by many months and sometimes years. This led to incompatibilities and errors at other sites.

It is proposed for the CMIP planning of data management to…

- motivate the ESGF's software forge to review and stabilize the software production processes. Hereby transparently prioritising the development of the different tools. This may include a central point where proposals for enhancements can be made,
- make sure that CMIP relevant data nodes install new software versions within an agreed, reasonable time, e.g. a month,
- make sure that a software release is accompanied by extensive debugging, and appoint a responsible person for debugging issues encountered after deployment. A test suite for automated ESGF debugging would be helpful[6].

3.5 Application of Software

In CMIP5 non-uniform and inconsistent application of software led to inconsistencies between data on hard disk (HD), metadata in data files, MD in the data node database, and between original data and their replica. Just three examples: Change of data on HD without publish/unpublish or without updating their checksum led to inconsistencies between MD in the node DB vs. data on HD. Data update by external tools after first writing led to inconsistencies between MD in the file header and the data in the same file. And the usual change of original data without versioning caused inconsistencies with replicas.

CMIP5 has demonstrated that various persons with appropriate permission do not hesitate to "correct" data directly on disk, causing inconsistencies between HD and DB. To identify errors of this kind, a software tool was proposed that would run regularly via cron job. However, this idea was subsequently discarded.

In addition, the ESGF software allows for different ways of publication and unpublication of data, which can be carried out in various steps. This results in data inhomogeneity by inconsistent software handling. Here detailed definitions of the different software workflows and instructions on how software should be applied might help. One can object that determinations of this kind should refer to the data product rather than to the way to produce it. However, on the base of a common software stack it will be possible to define its common application (i.e., the commands including the operands used). A detailed definition of the output data structures in projects of CMIP size would be much more difficult.

All in all, some partners seem not to have the resources to run their nodes[7] properly and as needed. We propose that ESGF partners encourage collaborations between different sites where those with fewer resources can be supported by others. We also propose that the CDNOT urge specific nodes to do so. This especially should be possible in funding communities like IS-ENES.

It is proposed for the CMIP planning of data management to…

- establish automated continuous consistency checks on data nodes of data and the describing MD. The necessary corrections need to be carried out,
- define in detail the application steps and parameters of the ESGF software,
- decide on basis of a statistics of usability and downtimes of the data nodes on support offered to them,
- offer workshops and/or tutorials to administrator users of the ESGF software (DN operators).

These recommendations have direct impact on the CMIP6 data management and should be covered by CDNOT as the CMIP6 data node operations team.

[1] WIP White Paper „CMIP6: file names and netCDF global attributes“ [https://docs.google.com/document/d/1kPQuKJyohCttdqZxzVldOqBUs94hsYer0YphBgWOZUc] with two additional documents of further specifications on CVs [VSPEC: https://docs.google.com/document/d/1CzTUoX4H2S0XbQUM3_9yKvJ2la7qUExFV7ibGzThmhA and VRECORDS: https://docs.google.com/spreadsheets/d/1ihUI7WbJrShTDUER862LBF2_4XQmxZiZMm_n5y1d-cs]

[2] English proverb referring to scientific data ;-) .

[3] NetCDF bears most MD essential for data use in the file header. This comprises variable name, unit, coordinate systems etc.

[4] WIP White Paper „CMIP6: file names and netCDF global attributes“ [https://docs.google.com/document/d/1kPQuKJyohCttdqZxzVldOqBUs94hsYer0YphBgWOZUc] with two additional documents of further specifications on CVs [VSPEC: https://docs.google.com/document/d/1CzTUoX4H2S0XbQUM3_9yKvJ2la7qUExFV7ibGzThmhA and VRECORDS: https://docs.google.com/spreadsheets/d/1ihUI7WbJrShTDUER862LBF2_4XQmxZiZMm_n5y1d-cs]

[5] WIP White Paper „Persistent Identifiers for CMIP6: Implementation plan“, Chapter 2, [https://docs.google.com/document/d/13VjI377yNRnBE9fHkAqRY5o-QlUQvlFaoTkPdwZaNio]

[6] The development of a test suite for ESGF debugging was started by IPSL (N. Carenton).

[7] This yields for Index Nodes, Identity Providing Nodes (IDN), Compute Nodes, and for data nodes. It is, however, most important for IDN.