Proposal: Scholarly citations for CMIP5 model output

Michael Lautenschlager, V. Balaji, Bryan Lawrence

March 11th, 2010

Summary

A key part of the CMIP5 process will be the expectation that data users will use citations to both credit those who produced simulations and identify unambiguously which data were used in any analysis.

This document proposes a method of doing this which leverages existing investments in data publication and documentation at the IPCC data distribution centres at the World Data Centre for Climate (WDCC) at DKRZ in Germany, at the British Atmospheric Data Centre (BADC), and at the Programme for Climate Diagnosis and Intercomparison (PCDMI) at Lawrence Livermore Lab in the U.S.

The methodology consists of a series of steps that will be jointly undertaken by the modelling centres submitting data to the archive, and the aforementioned data centres responsible for data distribution and persistence:

1. When data is published into the Earth System Grid Federation, all data will be quality controlled at the source, and some of the requested data replicated into the Federation Archival Centres.
2. Data which appears in the archival centres, and passes further quality control, will enter a scientific data publication process.
3. For entry into the archive centres, and to be considered for subsequent publication, all data must have adequate metadata entered via the CMIP5 metadata questionnaire.
4. The WDDC will be responsible for publication: assigning and registering persistent identifiers using the DOI system to certain versions of every model ensemble (whether it has one or multiple members) produced at every modelling centre for each CMIP5 experiment.
5. Once registered, the data will be persisted initially by the archival data centres within the Earth System Grid Federation, and for those data available for the IPCC fifth assessment report, indefinitely within the IPCC data centre federation – provided the data has been made available with appropriate access rights.
6. Once registered, a published data entity can uniquely be located by resolving its DOI and it possesses a specific citation reference.

The DOI will point to a page to be hosted at British Atmospheric Data Centre (and to be transitioned to the IPCC data distribution centre), which will include essential citation information along with appropriate metadata. Data access will be possible via links from that page into the Earth System Grid Federation Gateway at PCMDI.

Data compilations (such as multi-model ensembles) may also be created and lodged in the Earth System Grid Federation archival centres, and themselves enter other publication processes. Suitable publication processes would include the peer review processes of the Earth System Science Data journal (ESSD, http://www.earth-system-science-data.net/) and possibly Atmospheric Sciences Letters (under discussion with the Royal Meteorological Society). These journals would assign their own DOIs to point into the data holdings of the Earth System Grid Federation.

1. Introduction

Model output archives such as the IPCC and CMIP archives enable scientists to write papers based on runs done by others. We propose a method by which credit can properly be assigned for the teams that perform model runs. The intent is to create a publication of climate model data entities related to a defined citation for model output. The granularity of climate model data entities which is suitable for scientific literature seems to be on the level of experiments. This is the level of granularity which has been successfully implemented in a publication scheme by the World Data Centre for Climate (WDCC) at the DKRZ (German Climate Computing Centre)[1].

The WDCC scheme was developed within a consortium made up of scientific long-term data archives (WDC Climate, WDC-Mare, WDC-RSAT and GFZ-Potsdam) and the German National Library of Science and Technology (TIB), which developed the concept of 'Publication and Citation of Scientific Primary Data' (STD-DOI)[2]. The implementation is available at the TIB and first scientific data publications can be obtained from the TIB library catalogue together with classical scientific publications. More information in addition to the web page can be obtained from Brase (2004) and Klump et al. (2006).

In the remainder of this document we discuss in detail the way the STD-DOI scheme is implemented at WDCC before discussing how we believe a similar scheme should be implemented for CMIP5 and the IPCC.

2. Publication Process at WDCC

In this section we describe the existing publication process at WDCC, which we expect to use as a model for the IPCC/CMIP5 publication process.

The STD-DOI service requires several organizational and technical pre-conditions to make primary scientific data citable as publications:

- Quality control of the primary data set by the author and by the data publishing agency,
- Quality control of the descriptive metadata set by the author and by the data publishing agency
- Long-term availability of the published data in online repositories
- Search function for data publications in library catalogues (e.g. TIBORDER)
- Access to the primary data with assignment of a persistent identifier and resolver system (DOI resolver)
- Published data entities are fixed and can no longer change (subsequent versions may be published as later editions if appropriate).

This DOI assignment is seen to be integrated in a publication process and requires final quality control of data and metadata. As it is currently implemented, the main steps for scientific data publication at the WDC Climate are:

- Quality control of the scientific primary data and descriptive metadata by the authors and Publication Agencies (here WDC Climate)
- Creation of STD-DOI metadata (metadata for publication of electronic media)
- Creation of persistent identifiers (DOI/URN)
- STD-DOI metadata integration into library catalogue (TIBORDER)

1 World Data Center for Climate: http://www.wdc-climate.de

2 STD-DOI: http://www.std-doi.de

- Link to primary data and metadata at the WDCC archive level by integration of the persistent identifiers in the resolver systems (DOI resolver / URN) by the Registration Agency (here TIB).
- Primary data and metadata access via internet and graphical web-browser interface of WDCC.

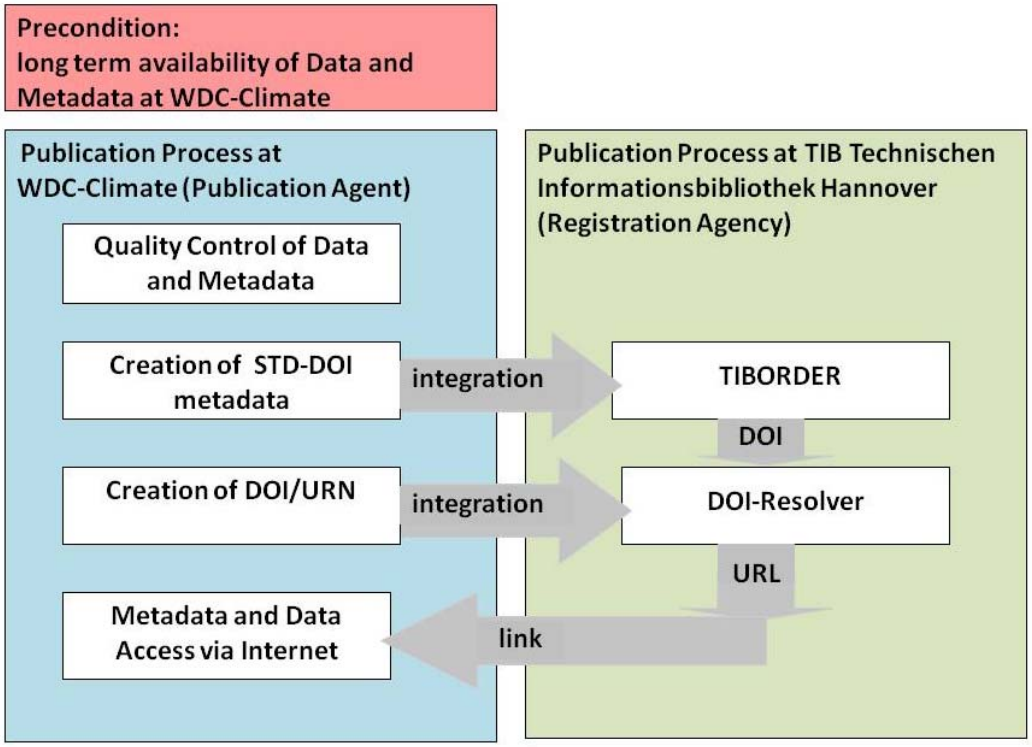


Figure 1: The STD-DOI publication process at WDCC.

The STD-DOI metadata integrate the standard DOI metadata and metadata according to ISO 690-2 which are suitable for electronic publication. The detailed definition of the STD-DOI metadata profile can be obtained from http://www.icdp-online.org/contenido/std-doi/upload/pdf/STD_metadata_kernel_v3.pdf and the corresponding XSD definition is located at http://www.tib.uni-hannover.de/doi/std-doi.xsd.

Data Quality Assurance at WDCC

The quality assurance procedure contains metadata and climate data. For climate data semantic and syntactic quality assurance are distinguished. Semantic and metadata quality assurance are performed in cooperation with the data provider and will be documented in the metadata as part of the provenance information.

The syntactic quality assurance is performed by the corresponding data archive. Technical Quality Control (TQA) of data are developed and implemented at WDCC for the publication of primary data together with persistent identifiers DOI and URN. WDCC's TQA ensures consistency between data and metadata within the WDCC's database tables (field-based data access). The check list and corresponding procedures are used for simulation data and are presently expanded to meteorological measurements:

1) Number of data sets is correct and not equal 0
2) Size of every data set is not equal 0
3) The data sets and corresponding metadata are all accessible via internet
   (this does not exclude a personal account and the data may restricted for free use in research

only.

4) The data size is controlled and correct
5) The time description (metadata) and existence of data are consistent. The data is complete and no duplicates exist corresponding to the metadata description
   a) start- stop date of metadata and header information are consistent
   b) Number of accessible data units is consistent with metadata description
   c) The continuous time steps (metadata) for the accessible data units are correct
   This means no vacancies or links exist in the metadata description of the accessible data units
6) The format is correct
7) Variable description and data are consistent.

Quality assurance of the content of climate model data is mainly in the responsibility of the data authors (SQA – Semantic Quality Assurance). Quality assurance is documented in the metadata and the assigned quality flag in the database system is "approved by author".

Citation of Data at WDCC

The size of the data sets used in a scientific publication often prohibits their publication as data tables and, as a result, data used as the basis of a publication are rarely published any more. The lack of access to scientific data is an obstacle to interdisciplinary and international research.

Persistent identifiers together with their bibliographical information provide the opportunity to find and to cite primary data in scientific publications.

A citation of a data set follows the classical citation rules in scientific literature, e.g. author(s), publication year, data set name, persistent identifier.

Examples of data citations

Nozawa, Toru (2004): IPCC-DDC_CCSRNIES_SRES_B2: 211 YEARS MONTHLY MEANS, National Institute for Environmental Studies and Center for Climate System Research Japan, *WDCC*. doi:10.1594/WDCC/CCSRNIES_SRES_B2

Kamm,H; Machon, L; Donner, S (2004): Gas Chromatography (KTB Field Lab), *GFZ Potsdam*. doi:10.1594/GFZ/ICDP/KTB/ktb-geoch-gaschr-p

Stein, R.; Fahl, K. (2003): Distribution of grain size and clay minerals in surface sediments of the Kara Sea, *PANGAEA*, doi:10.1594/PANGAEA.119754.

Application in the literature

Lorenz, S.J., Kasang, D., Lohmann, G. (2005): Globaler Wasserkreislauf und Klimaänderungen - eine Wechselbeziehung, **In:** *Warnsignal Klima: Genug Wasser für alle?* Lozán, Graßl, Hupfer, Menzel, Schönwiese (Eds.), pp. 153-158. Wissenschaftliche Auswertungen, Hamburg, Germany.

This article uses and cites:
Stendel, M., T. Smith, E. Roeckner, U. Cubasch (2004): ECHAM4_OPYC_SRES_A2: 110 years coupled A2 run 6H values. *World Data Center for Climate*. doi:10.1594/WDCC/EH4_OPYC_SRES_A2.

Granularity at WDCC

The complete list of data entities which are published in the TIB library catalogue via STD-DOI can be obtained from TIBORDER (http://tiborder.gbv.de/DB=2.63/LNG=EN/) choosing WDCC as search item. The result is a list of 114 data entities which ran through the WDCC publication process and which are integrated as WDCC data publications in the TIB library catalogue together

with scientific literature entries. The total amount of STD-DOI data publications counts more than 1800 entities from different scientific disciplines.

WDCC defines independent data entities for climate model data at the level of individual model experiments.

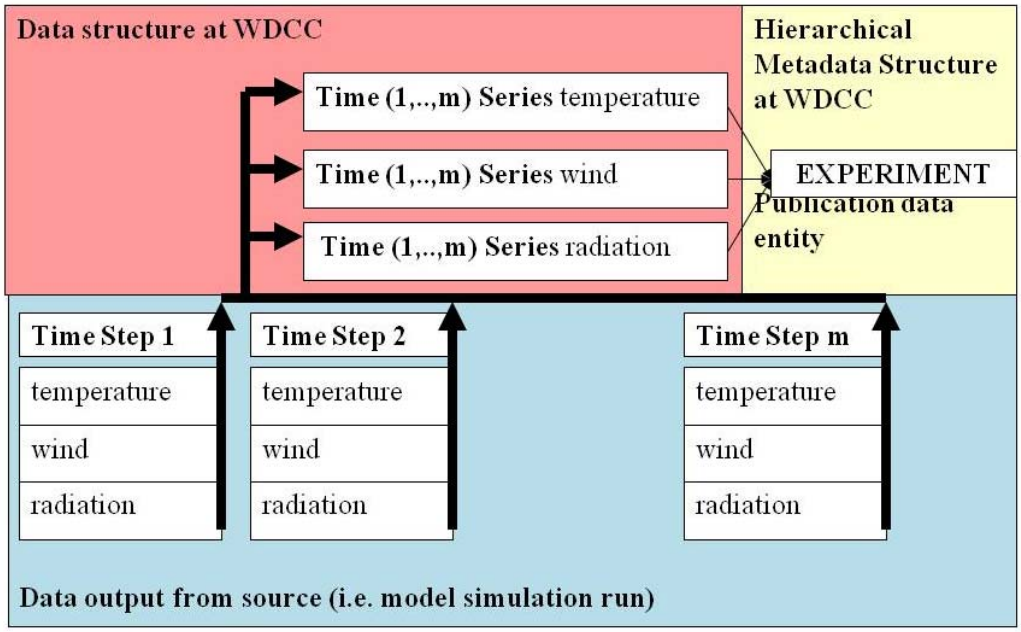


Figure 2: The relationship between data sources and data structures at WDCC. It can be seen that the WDCC time series (data base tables) correspond with atomic data sets in the CMIP5/AR5 environment.

Essentials of WDCC's scientific data publication service

The STD-DOI publication process (Details: http://www.mad.zmaw.de/projects-at-md/publication-and-citation/) includes assignment of persistent identifiers (DOI and URN) and of citation directive together with data entity integration in the TIB library catalogue for interdisciplinary data usage. The STD-DOI data publication process is part of the long-term archive implementation of WDCC and DKRZ. Important aspects of WDCC's scientific data publication service are in summary:

- The identification of independent data entities which are suitable for publication at the complexity level of scientific literature reference lists,
- The execution of an elaborated review process for metadata and climate data (quality control),
- The assignment of additional metadata for electronic publication (ISO 690-2) and of persistent identifiers (DOI / URN) and
- The integration of publication metadata and persistent identifiers into the TIBORDER library catalogue (German National Library of Science and Technology, Hannover) so that primary data entities are searchable and citable together with scientific literature.
- Quality characteristic is presently "approved by author", could be "peer reviewed" with ESSD (Earth System Science Data Journal).
- Published data entities cannot be modified any longer.
- Data are freely available via Internet.

The STD-DOI data publication at WDCC is ready to use and it is suggested to use it for assignment of persistent identifiers and citation directives to CMIP5/AR5 data climate model data entities. The STD-DOI publication process requires suitable metadata even at the scientific level. These metadata are closely related to the METAFOR metadata questionnaire and the resulting CIM repository. The envisaged CMIP5 data and metadata quality checks can be directly connected with the ESG data publication process. DOI assignment is closely related to versioning. Each new version requires a new persistent identifier and citation directive in order to identify the data entity precisely within the literature.

3. Implementation of the CMIP5/AR5 model data publication process

A flow diagram of the CMIP5 data ingest and publication together with tasks for data acceptance and publication is provided in Figure 3.

The CMIP5/AR5 data acceptance and publication is mainly related to three activities: ingest control of data and metadata, generation of the CIM data repository and the STD-DOI data publication process.

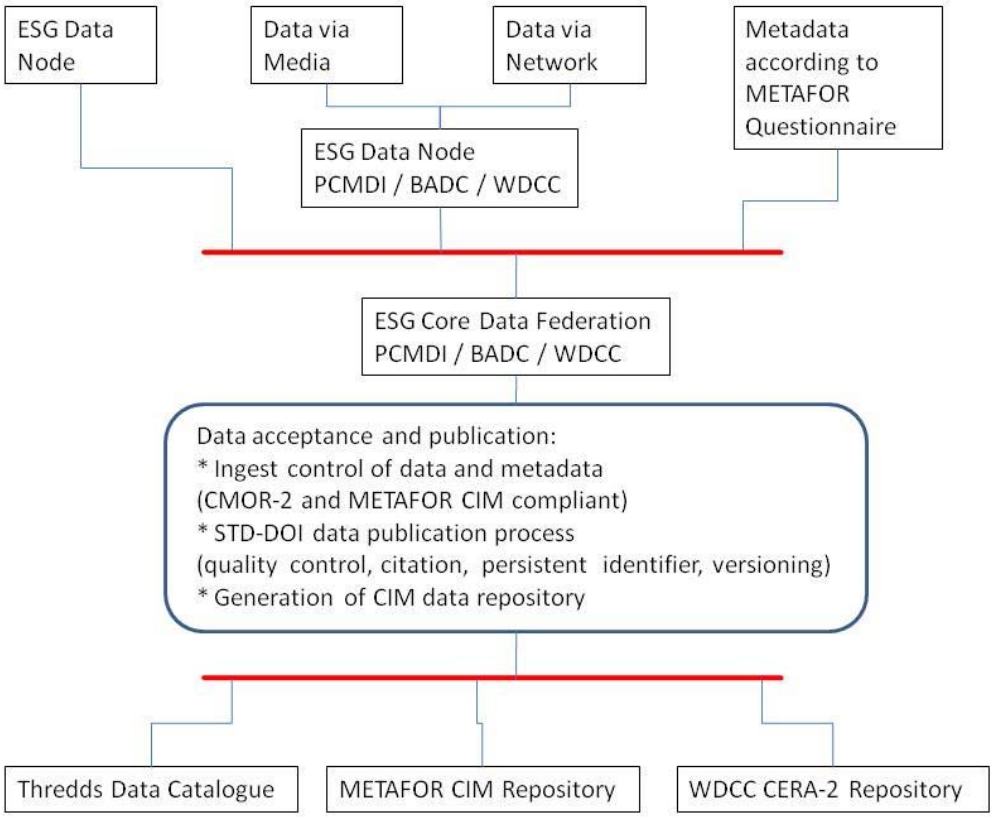
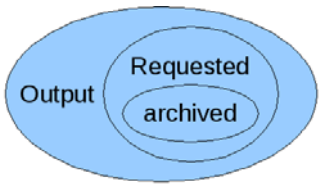


Figure 3: In this figure, the bottom red line indicates data which has passed all quality control, and is accepted into the archival centres, and into the publication process. (Thredds data catalogues will also expose data at the originating model centres.)

If we start by considering the ingestion process, we recognise a number of phases of ingestion and publication within ESG:

1. Modelling groups load data onto "their" ESG data node (or send data to a remote site to host their ESG data node, but in what follows that remote copy is still "their" ESG data node).
2. The ESG data node carries out an ESG data publication process, which makes the data visible to ESG gateways. Not all that data will be "CMIP5 requested data" and not all the requested data is intended to be replicated into the "CMIP5 archive centres" (Figure 4):

3. Data destined for the Federation archives will be replicated to PCMDI, BADC, DKRZ and others.
4. Data replicated to the archive centres will itself enter an ESG publication process to make these data available to ESG gateways.
5. Updated versions will follow steps 1 to 4 as appropriate.
6. Data held in in the archival centres will meet the STD-DOI requirements for long term persistence, will transition into the IPCC Data Distribution Centre, and are thus suitable to enter a DOI publication process.



Output: by modelling centre
Requested: by CMIP5
Archived: by PCMDI, BADC, DKRZ and others

Figure 4: Data output by modelling centres will exceed that requested by CMIP5, and only some of the requested data (approximately half) will be archived (and replicated globally).

Thus we suggest the CMIP5/AR5 DOI publication process should exploit and extend the ESG procedures by modifying step 4 so that we have a synchronised publication procedure which yields the publication of archived datasets both into ESG gateways, and into a DOI assignment process.

This synchronized publication procedure yields some benefits and would result in the following:

- The CMIP5/AR5 data federation provides quality proven data.
- Each model data version is connected to an agreed citation direction.
- Each model data version can be identified and accessed by an individual persistent identifier (DOI/URN).
- The STD-DOI scientific data publication allows for model data search together with scientific journal publications in library catalogues (TIBORDER).
- The STD-DOI scientific data publications allows for data access directly by DOI and the IDF global handle server independent from the actual archive location.
- The Metafor portal can be used to display the citation information and launch users to the underlying data. (The Metafor infrastructure could also be used to ensure that users can take an individual file metadata, and go direct to the appropriate citation information, making it easier for users to cite data appropriately.)

But the STD-DOI data publication is connected to a few requirements:

- The more elaborate quality assurance of the STD-DOI data publication must be finished before the final ESG data publication (in step 4) in order to ensure consistent versioning, referencing and citation.
- Data are fixed and no longer matter of change after the STD-DOI publication and persistent identifier assignment. Small errors can be addressed as “Erratum”, larger modifications result in a new version of the data entity with a new persistent identifier (DOI/URN).
- The existing STD-DOI scientific data publication process requires metadata data entry and data integration in the WDCC data repository CERA-2.
- Quality assurance must be designed to work with 2 or 3 three times of the amount of replicated CMIP5/AR5 data within the expected period of data acquisition (approximately one year); it is our experience that model data normally has to be sent more than once because of errors. We will also need to support metadata entry.

STD-DOI Quality Assurance

Ideally the STD-DOI related quality assurance will be identical to that quality assurance procedure of the ESG+IPCC core data archive federation which itself is related to the ESG data publication process as discussed above. The same quality assurance for the STD-DOI scientific data publication and for the ESG data publication is not necessary but it is desirable in order to use synergy effects and to save work during data archive population.

In order to establish a clear and accepted quality flag for the ESG core data archive entities it is essential to implement an agreed, well documented and transparent quality assurance process. This quality assurance process is discussed in a parallel paper "The CMIP5/AR5 Model Data Quality Control". Explicit in that document are three levels of quality control:

- Level 1: CMIP5 model results passed CMOR-2 conformance checks and ESG data node conformance checks. All data available via the ESG consortium gateways is expected to meet this level of quality control. Model metadata may or may not have been produced.
- Level 2: (Can only apply to requested data replicated to archive centres). Model metadata is available. Additional automatic checks have been completed, along with subjective checks of samples of data and metadata.
- Level 3: (Can only apply to requested data replicated to archive centres). The STD-DOI publication validation has been passed.

Integration of STD-DOI data publication into the ESG data publication process

At the end of the ESG data and metadata publication process in the archive nodes (step 4 above) we assume everyhing has passed the CMIP5-ESG quality control (discussed below): CMOR-2 compliant data are available after the ESG publication process and in parallel CIM compliant metadata are available from the METAFOR CIM repository (compare process diagrams from CMIP5-ESG quality control).

At this point the STD-DOI publication process starts by requesting a DOI for the ESG published data entity of the CMIP5/IPCC core variables. These data have passed first quality control steps but they are still matter of change. From the point of view of the STD-DOI publication old data could be replaced by new versions without keeping the old ones and data access could be restricted. This will be changed after finalization of the STD.DOI publication process and DOI assignment. (Persistent identifiers are related to persistent objects.) After final DOI assignment data are no longer matter of change and published scientific data must be accessible at least throughout the scientific community.

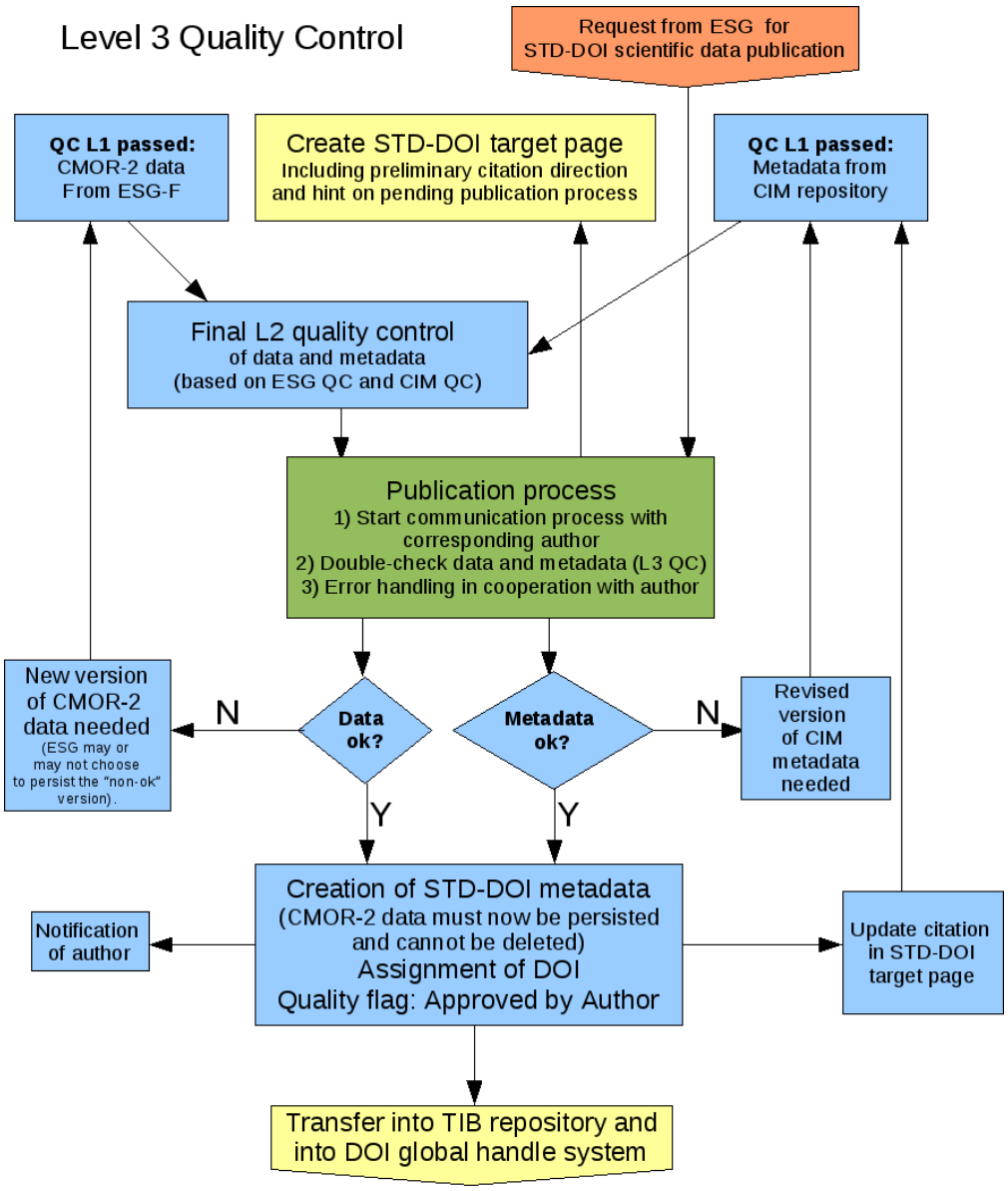


Figure 5: The DOI publication process and relationships to other levels of quality control within the ESG system.

STD-DOI publication process steps:

1. Start data publication with creation of STD-DOI target URL which is hosted at BADC. The target URL is an HTML page which provides the basic data entity information after resolving the DOI in the TIB library catalogue. Basic information at that page is Data citation direction:
    1. Authors (publication year): Title of data entity including institute/modelling group. World data Center for Climate. doi:10.1594/WDCC/activity_institute_model_experiment_frequency_vxxxxx (The DRS hierarchy together with the version number is unique for the simulation, version yet not specified)
    2. Contact for data entity
    3. Link of referenced data entity to entry in CIM repository at BADC
    4. Link of referenced data entity to primary data access point at PCMDI with identification of mirror copies in core data archives

2. Quality Control of data and metadata with respect to completeness, correctness and consistency. Ideally the the ESG quality assurance process is complete with respect to STD-DOI data publication requirements. Then the log-file of the ESG quality control will be double-checked and no additional quality assurance will be necessary.
3. Publication Process itself:
    1. Integration of metadata into WDCC repository to accomplish the data publication process
    2. Double-check with corresponding author data and metadata entries ,
    3. solving open questions,
    4. updating data and metadata, (updating the model data requires a new ESG ingestion and publication process; updating metadata goes directly into the CIM data repository, a new questionnaire dissemination process will normally not be necessary)
    5. approval from author to freeze this data entity (quality flag is “approval by author”)
4. Creation of STD-DOI metadata, assignment of DOI (including frozen ESG data system version number), ingestion into TIB catalogue and DOI handle system and updating the STD-DOI target page with final DOI (completed by version to be published “doi:10.1594/WDCC/activity_....._version”)
5. Notification of author about finalization of data publication process. At this point data are globally accessible by the DOI and generally open for scientific use.

The complete process is shown in Figure 5.

Publication Time-scales and Data Availability

(This section will need WGCM approval, so both Karl and Ron need to buy agree this before it can even go to WGCM … and it may even need TGICA approval).

The various levels of quality control will take time, and the publication process itself will also take time. It has been agreed that data which has passed Level 1 quality control will be available to the community immediately, but given that at this stage the data has not been quality controlled to publication standards, we recommend that this data be only available to the modelling community, who will be best placed to use these data reliabily – and contribute to the subjective evaluation necessary for the second level of quality control. We recommend that the access constraints to the data during this period be the same as those used for CMIP3 prior to AR4 (essentially restricted to what is notionally the IPCC WG1 community).

Data which has passed the second level of quality control will be available from PCMDI, BADC, DKRZ and others (as well as the originating data nodes). At this point it will have entered the STD-DOI publication process, but this could take some weeks as we expect a considerable amount of data to arrive for evaluation in a short period of time.

When the data has passed level 3, the data has been formally published, and we would expect the access constraints to be the same as any academic published entity: freely available subject to copyright and the expectation that use will result in citation. Data which has passed to level 3, and which meets the same timing criteria as used by the IPCC working groups for “normal” publications, will be persistent in the IPCC data distribution centres. Modelling centres who cannot subscribe to this level of openness could not have their data enter the STD-DOI publication process, nor have their data persisted in the IPCC archive.

During the publication period, scientific articles must cite these data according to the preliminary data citation reference which is given in the DOI target URL. The DOI target URL will be established at the begin of the STD-DOI publication process and will be hosted at BADC together

with the METAFOR CIM data repository. Before final acceptance of the scientific paper the preliminary data citation reference must be replaced by the finalized DOI version (also hosted at BADC).

During the STD-DOI publication process scientist must be aware that subsequent versions of the dat may be created, and while the ESG system should notify them of new data, they should not rely on it.

Granularity

An additional aspect in the CMIP5/AR5 data publication is the specification of the granularity of independent data entities which is suitable for scientific journal publications. The granularity must be fine enough for direct data access and it must be coarse enough not to overburden reference lists in scientific journals. Data citations and citations of scientific articles must be balanced in the reference lists in order to achieve acceptance of scientific journal publishers and editors as well as of readers.

The currently implemented STD-DOI scientific data publication service at WDCC/DKRZ is based on individual model simulations which summarize all numerical model data which are calculated for one experiment. All ensemble members are summarized under one DOI.

Following the same procedure for CMIP5/IPCC would result in 1000 and 1500 DOIs in the direct STD-DOI publication for the core data archive. These DOIs would allow for direct data access to one specific simulation from one climate model. The DOIs do not allow for direct data access at the granularity of ensemble members or runs or at the level of specific atomic data sets (variables etc). Access at this level of granularity will be permitted by the DOI pointing via the METAFOR portal to underlying scientific data and metadata systems (such as the ESG gateway itself).

Beside this basic STD-DOI publication process new data collections and data products which are created during the CMIP5/IPCC data evaluation process can be published additionally, either using the same methodology, or by other academic journals such as the Earth Science Data Journal (http://www.earth-system-science-data.net/).

REFERENCES

Brase, J. (2004): Using Digital Library Techniques - Registration of Scientific Primary Data. Lecture Notes in Computer Science 3232, 488-494.

Guilyardi, E., 2006: El Niño: mean state-seasonal cycle interactions in a multi-model ensemble. *Climate Dynamics*, **26 (4)**, 329–348.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I. and Wächter, J. (2006): Data publication in the Open Access Initiative. Data Science Journal 5, 79-83. doi:10.2481/dsj.5.79.