CMIP6 Data Citation and Long-Term Archival

Authors: Martina Stockhause, Frank Toussaint, Michael Lautenschlager

Date: 2015-08-05
Version: 3bnl (Exec summary and objective modified by Bryan)
Version: 4 by authors (Split of document: section 4 on quality was moved to a separate document; section 6 on implementation was revised and integrated in section 4 together with the new workflow description)
Version: 5 by authors (Modifications agreed on WIP telco on 2015-08-04)

Scope

This document will discuss/develop and promote a consistent data citation regulation for CMIP6 data.
It will cover the description of the different data citation paths used for early citations after ESGF data publication towards the citation of stable reference data in the IPCC DDC AR6. The requirements for citations will be laid out in respect to Force 11’s Joint Declaration of Data Citation Principles (http://www.force11.org/datacitation) and including technical needs within the ESGF data infrastructure as well as data policies.

Content

Executive Summary

For the dynamic CMIP6 data disseminated by ESGF an early citation reference is requested in close connection to the ESGF data publication process. This citation reference have to meet the requirements of Force 11's 'Joint Declaration of Data Citation Principles' including the availability of citation information (e.g. authors and title) and a Persistent Identifier (PID) for data and data documentation access. The use of HDL Identifiers (CNRI Handle System Identifiers) is planned. For the CMIP6 data subset transferred in a long-term archive (e.g. in the IPCC DDC AR6), the established DataCite data publication using DOIs is foreseen. The DataCite citations will be seamlessly related to the (early) citations of the dynamic CMIP6 data via PID relations and vice versa. The data citations are offered on model and simulation granularities to meet different citation requirements in literature. The main requirements are related to technical as well as organizational aspects, including:

- Version support for the identification of the cited data collection,
- Integration of non-ESGF information into technical infrastructure for CMIP6, e.g. ES-DOC, QC,..., for the enrichment of data documentation in the long-term archives,
- Commitment to and supervision of data management agreements outlined in a Data Management Plan (DMP) for data consistency across ESGF data nodes, and
- Overall operability of the technical infrastructure.

Specific Implications for the Different Actors

ESGF / Technical CMIP6 Tasks:

Requirements of data citations for the ESGF are needed in two basic functionalities: version support and integration/support for external information like CIM or Errata. These affect the CoG front end, the search API and the replication (**Table 1**).

Table 1: Overview over data citation requirements for the ESGF

ESGF component	Requirement	Section	What for in Data Citation / LTA	Importance
CoG frontend	Integrate link to citation information for datasets using template link.	4.7	Dynamic CMIP6 data citation accessible for ESGF users	high
CoG frontend	Persistent HTML link to data collection on citation entry for use case: "All datasets in all available versions for a search request based on facets."	4.2.3	Data Access link on HDL Identifier landing page for dynamic CMIP6 data	high
Search API / Version support	Reduce the dataset result list for use case: "All datasets in the latest version for a specific date in the past".	4.4	Identify datasets in those versions, which were available at the cited access date for a citation entity.	high
Search API /	Use case: "Show differences	4.4	Which datasets of the citation	medium

Version support	of the results for a search query latest dataset versions for a specific date in the past to the results for a current date search query."		entity have been revised since the access date.	
Version support / Errata service	Use case: "Show newer available version of a given revised dataset and show errata information."	4.4	Why have datasets been revised? Scientific difference between cited dataset version and current dataset version.	medium
Replication	Use case: "Replicate latest version for a fixed/past date."	4.5	LTA/IPCC DDC	high
Replication	Use case: "Replicate subset specified by facets"	4.5	LTA/IPCC DDC	medium
Replication	Use case: "Replicate available related information together with the data: CIM, Errata, QC,…"	4.5	LTA: enrich metadata for non-expert data users	medium
Controlled/ Standard vocabulary for DRS components	consistent use of DRS components	4.3	Connect dynamic citations to CMIP6 infrastructure (ESGF, ES-DOC, LTA); technical quality assurance for LTA	high
ES-DOC/CIM	Use case: "Look-up possibility for citation information from the file header without going back to the portal"	4.7	Integration of citation information in furtherInfoUrl target page; furtherInfoUrl is a global attribute in the file headers	high

CDNOT Tasks:

CDNOT is expected to co-ordinate the ESGF implementation of the data citation and LTA requirements and to supervise the DMP compliance as well as the operability of the technical infrastructure (ESGF, ES-DOC, and other systems). The supervision of DMP compliance includes the additional tasks of *'QA data nodes'* (data nodes selected by WIP), which check the availability and quality of the citation information delivered by the modeling centers.
A web-accessible list of ESGF data node contacts is provided with information about, whether it is a 'QA data node' or not and which modeling centers an individual 'QA data node' is responsible for.

Modeling Center/Data Creator Tasks:

Modeling Centers are expected to provide data citation information their data on the granularities:

- ***Citation of model data:*** all datasets provided for CMIP6 by a single model, and

- ***Citation of simulation data:*** all datasets provided for a CMIP6 experiment (all ensemble simulations).

If no citation information for simulation data is provided, the citation information for the model data is used. The delivery of personal PIDs like ORCID for data creators/authors is recommended.
Initial citation information for model data including a contact person is collected prior to ESGF data publication together with other information, coordinated by the WIP (see workflow in section 4.1).
Citation information can be completed and changed during CMIP6 using the Citation GUI. A data subset will be moved into the IPCC DDC and assigned DataCite DOIs on the same granularities (model and simulation data). The modeling center will be asked for approval. Afterward the citation information will be fixed.
The benefit for the modeling centers is to receive credit for their data by data users citing the data in their scientific publications.

Data User Tasks:

- ***Citation of Dynamic CMIP6 data in scientific publications as:***
  data creators (publication year):Title. Version. Publisher. HDL Identifier.
  Citation information is accessible in the ESGF frontend via the "data citation" button for a dataset.
- ***Citation of stable IPCC-DDC AR6 data in scientific publications as:***
  data creators (publication year):Title. DOI Publisher. DataCite DOI.
  Citation information is accessible in the IPCC-DDC and ESGF portals.

Two citation granularities are offered: citation of model data and citation of simulation data. As the number of data citations in a reference list has to be in balance with article citations, it is recommended to use model data citations for the intercomparison of multiple models and multiple experiments, and simulation data citations for an analysis of few experiments from one or multiple models (section 4.2.2). Additional data citation policies (terms of use, license information) will be available on the landing pages for DOI and HDL Identifiers.

1. Introduction

1.1 Objective

As the citation of data gets more prevalent, and scientific publication begin to have data citation requirements, CMIP6 needs to define rules and methods for how to cite the data at the very early stages of the scientific evaluation process. In doing so, they need to distinguish between the necessity to give credit to the data creators and enabling scientists and reviewers to reproduce or at least check scientific results.

A data citation consists of creators[1], the title, a publication year, and the data publisher. It includes a persistent identifier (PID) that facilitates access to the data and to information on the data. Examples for PIDs are DOIs and HDL Identifiers. DataCite DOIs additionally register citation metadata and some additional information like a short description (abstract) for the DOI referenced data entity.

Other PID systems for the identification of entities are e.g. DOIs for scientific publications or PIDs for researchers like ORCID. Rules for the interoperability of these different PID systems to form a research environment are currently investigated.

As many variables are delivered by each modeling group, i.e. data creating institute, data need to be collected before the granularity is suitable to be used as entry in a reference list. The ***granularity for data citations*** is coarser than the granularity for data access at the scientific working level. Within CMIP5 each simulation of a certain model was a citation entity.

The experience of CMIP5 has shown that parts of the data are revised several times before the data has overall stability. There are two levels of stability that need to be considered: dynamic data, which needs to be identified, but may change, and two classes of final data: stable data which may not be long-term archived (and may be retracted in future), and stable data which will be long-term archived (in, for example, the IPCC DDC for long-term interdisciplinary use). That means that during the project phase (of CMIP6) parts of an early cited data entity in ESGF may change over time (dynamic data) or vanish whereas cited long-term archived data entities are stable and available for the long-term data reuse. Thus two interconnected citation mechanisms are required, the early CMIP6 citation reference of dynamic data and the DataCite IPCC-AR6 citation reference of stable data.

The citation of CMIP6 data is closely related to:

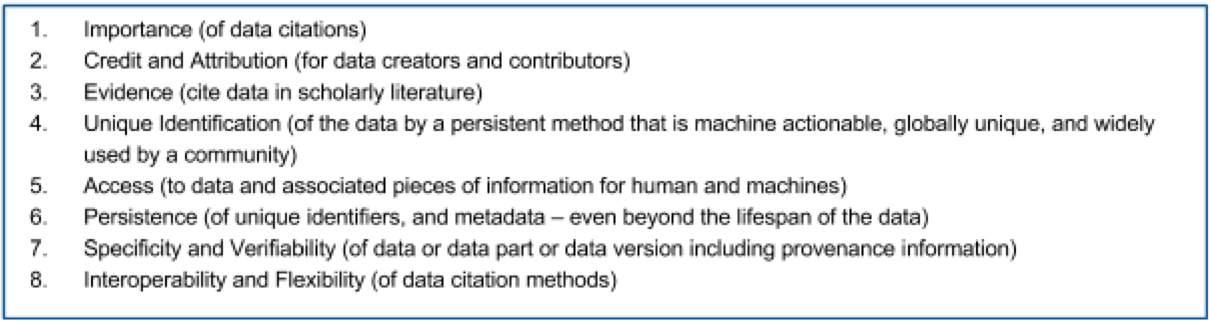
- a consistent and consequent ***versioning policy for data consistency and verifiability*** together with services for producers and consumers as precondition (WIP versioning white paper),
- the compliance of the data to certain quality standards (WIP quality white paper), and
- the availability of comprehensive information on the cited data in respect to the creation and provenance process (CIM documents, errata information) and data quality.

The compliance to WIP specifications including those for data citation should be mandated by the WIP and supervised by the CDNOT. Specific thoroughness should be applied to the quality assurance of the CMIP6 ***Core data*** subset which is aimed for long-term persistence in the IPCC DDC AR6 data.

[1] Creators are lists of persons or institutes in priority order. According to the DataCite definition, creators are main researchers or research institutions/groups involved in producing the data, or the authors of the publication. Editors might be credited as creators.

1.2 Background

As a broad international consensus Force11 (The Future of Research Communications and e-Scholarship) formulated a ***'Joint Declaration of Data Citation Principles'*** consisting of the 8 principles:



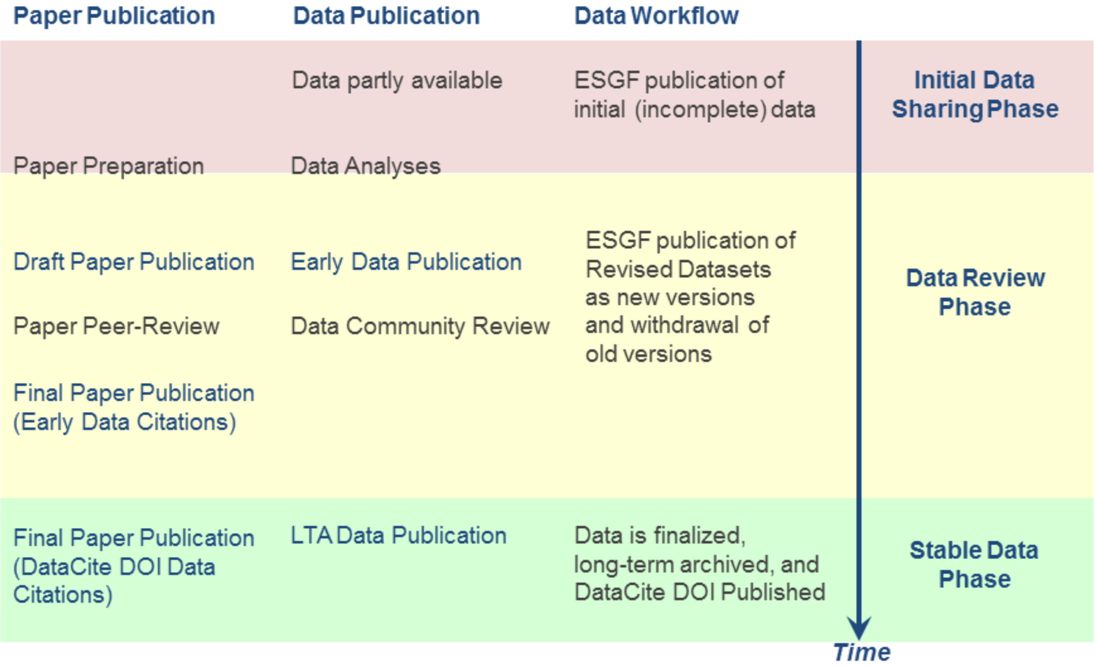
1. Importance (of data citations)
2. Credit and Attribution (for data creators and contributors)
3. Evidence (cite data in scholarly literature)
4. Unique Identification (of the data by a persistent method that is machine actionable, globally unique, and widely used by a community)
5. Access (to data and associated pieces of information for human and machines)
6. Persistence (of unique identifiers, and metadata – even beyond the lifespan of the data)
7. Specificity and Verifiability (of data or data part or data version including provenance information)
8. Interoperability and Flexibility (of data citation methods)

These minimum requirements for the citability of data have to be fulfilled by the data citation approach for CMIP6/IPCC DDC AR6. Principles 4 to 7 aims at making cited data persistently (re-)accessible by its persistent and unique identifier (PID). If data is no longer available, the PID target should provide at least information on the data ("tombstone" web page). For the dynamic CMIP6 data that (esp. principle 7) requires consistent and strict versioning, and a possibility to cite versions of data collections. The data might be no longer available in a certain version as it gets revised and published in the ESGF under a new version, but the information on its previous version remains (a PID on such a file should point to a tombstone page). Ideally, the PID target page for the old and unpublished data version should include errata information and provide a link to the latest (revised) data version.

The Data Publication Model of CMIP6 is the ***Standalone Data Publication*** as described and classified by Lawrence et al. (2011; doi:10.2218/ijdc.v6i2.205), which means that the data publication is not directly coupled to a paper publication by the same authors but is a scholarly publication entity of its own. The ***data peer review process*** to assess and ensure the quality of datasets is discussed for different Data Publication Models in Meyernik et al. (2014, doi:10.1175/BAMS-D-13-00083.1).

Within the life cycle of the CMIP6 data three main phases for data and data citations can be distinguished as proposed by Stockhause et al. (2014, EGU2014-3266_presentation.pdf; Figure 1):

- *Initial Data Sharing Phase:* First dataset versions are available in the ESGF for download. Datasets might be added to a data citation entity (data collection). Data is shown in presentations to a selected audience or within the CMIP6 project.
- *Data Review Phase:* As the climate community starts to analyze the data, individual datasets are 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 publication).
- *Stable Data Phase:* Towards the end of the project data gets stable and long-term archived (LTA) for long-term interdisciplinary use by the IPCC DDC AR6 users (LTA data publications with DataCite DOI minting).



Paper Publication	Data Publication	Data Workflow	
	Data partly available	ESGF publication of initial (incomplete) data	Initial Data Sharing Phase
Paper Preparation	Data Analyses		
Draft Paper Publication	Early Data Publication	ESGF publication of Revised Datasets as new versions and withdrawal of old versions	Data Review Phase
Paper Peer-Review	Data Community Review		
Final Paper Publication (Early Data Citations)			
Final Paper Publication (DataCite DOI Data Citations)	LTA Data Publication	Data is finalized, long-term archived, and DataCite DOI Published	Stable Data Phase

Figure 1: Phases of the CMIP data workflow with respect to data and paper publications (according to Stockhause et al., 2014).

References:

- Force11 - Joint Declaration of Data Citation Principles' (http://www.force11.org/datacitation)
- DataCite Principles (http://www.datacite.org/whatisdatacite)
- ORCID (PIDs for researchers: http://orcid.org/) and ODIN (ORCID and DataCite Interoperability Network: http://odin-project.eu/)
- RDA/WDS Interest Group Publishing Data and related Working Groups (http://rd-alliance.org/group/rdawds-publishing-data-ig.html)

2. Types of Data Citations

It is always possible to informally cite data as long as information on data creators and other contributors and the title are available. Formal citations require additional reliable access to the cited data portion. As granularity for a data citation entity data collections of a simulation (all realizations) as in CMIP5 are proposed. The use of ORCID (PIDs for researchers) in the collected citation metadata is recommended in addition to creator names to enable creator name verification.

2.1 CMIP6 Citation of Dynamic Data (Incomplete and Changing Data)

The CMIP6 data belonging to a citation entity is dynamic, i.e. the content of the data collection identified/referenced by a PID changes over time. Therefore the information of the cited data version or access date needs to be added to a citation to identify the data and fulfill the specificity and verifiability principle. For consistent data versioning the latest version over all datasets should be used. For transparency reasons, a version string including the date is recommended.

We plan the use of HDL identifiers as PIDs in these citations maintained centrally by the citation repository. The reasons not to use DOIs are: Data accessibility or even maintenance of tombstone pages cannot be granted for all data nodes of the Earth System Grid Federation hosting parts of the data. Data nodes might be down for longer time periods or even switched off completely. Only data node managers can grant long-term access to the data but not every data node center will be a DOI publication agency.

Citation of Dynamic CMIP6 Data:

data creator list (publication year): Title. Version. Publisher. HDL Identifier.

2.2 IPCC-AR6 Citation of Stable Data (Complete and Unchanging Data)

For the finalized and unchanging data in the long-term archive of the IPCC-DDC hosted by WDCC the registration of DataCite DOIs is planned as for IPCC-AR5 data.

As only one data version is archived, the specification of a version in the citation is not required. To enable long-term interdisciplinary data reuse, comprehensive information on the creation process (CIM documents, errata information) and data quality is added to the archived data as accessible.

Publishers of scientific journals often have stricter requirements on data citations than Force 11, e.g. data completeness, persistence, curation or accessibility, which DataCite DOI publishers fulfill.

Citation of Stable IPCC-AR6 Data:

data creator list (publication year): Title. DOI Publisher. DataCite DOI.

2.3 Option for Data Publication in a Data Journal (Peer-Reviewed Data)

The CMIP6 quality assurance process (WIP quality paper) does not include a formal scientific data review process. An informal data review is performed during the Data Sharing Phase by scientists analyzing the data.

Peer-reviews on data are provided by data journals like ESSD (Earth System Science Data). A data paper is submitted with formal data citation entries in the reference list. For CMIP6 a PID to the CIM documentation should be added in the reference list. For CMIP6 a special issue might be initiated by WGCM.

Since a long-term availability is required, ***currently only IPCC-DDC*** data fulfills the data journal requirements of long-term data availability, permanent data accessibility, and data stability (finalized and stable data; see section 3).

References:

HDL Identifier: http://handle.net/factsheet.html.
DataCite Business Model: doi:10.5438/0007.

3. Citation Concept with respect to Data Citation Principles and CMIP6 Requirements

Citation and workflow policies including implementation aspects for the requirements in this section are specified in section 4. The described requirements are summarized at the end of this section in Table 2.

3.1 Credit and Attribution via Citation Information

The credit for the data creators and contributors[2] require the availability of those names as soon as CMIP6 data is accessible by users of the ESGF. For cross-checking of names, the use of ORCID IDs is recommended.

3.2 Unique and Persistent Identification, Access and Persistence

For unique and persistent identification, PIDs are required. The persistence of PIDs requires in general the persistence of the target URL as well as the persistence of metadata and its disposition. It does not necessarily include the persistence of the data according to Force11. If the data is retracted, the information about the data remains and preferably provenance information about the cause of retraction or about an available revised data version is added.
DataCite and the data journals (ESSD, GDJ, and Scientific Data) require additionally the long-term availability of data and describe the data as complete and not liable to change. They define standards for repositories or even recommend trusted repositories. GDJ require PIDs of type DOI on the referenced data. Long-term availability of CMIP6 data is probably restricted to the reference data as part of the IPCC DDC.

3.3 Specificity and Verifiability

The idea of formal data citations is, - apart from the credit given to the data creators -, the possibility to facilitate the verification of the content of a paper and thus the re-accessibility of the same data portion as underlying the paper and sufficient documentation on data content and provenance for interdisciplinary data usage. ***A strict and consequent data versioning is crucial to fulfill this criterion.*** The collection of errata information explaining data changes between versions is recommended.

3.4 Data Aggregation / Granularity

Apart from the general data citation principles of Force11, the granularity of a data citation must be suitable for use in reference lists of scientific publications (balanced to paper citations) to be accepted by the journal publishers. The data citation entity for data has to be a collection of jointly citable datasets. Subsetting of the data citation granularity (collection of data access entities) cannot be reflected in the citation reference. This has to be performed in the scientific paper itself.
Citations for CMIP6 model data as well as CMIP6 simulation data are planned to serve different user requirements on data citation granularity. The PID for such a data collection has to point to a target URL with access to all datasets belonging to this data collection. At least for the (early) CMIP6

[2] Data creators appear in the citation information, contributors do not. In comparison to articles creators can be viewed as authors and contributors as persons or institutes in an acknowledgement. Examples for DataCite contributor types are ContactPerson, Distributor, Editor, Funder, HostingInstitution, Researcher, and RightsHolder.

citation reference of dynamic data two granularity levels are envisaged, the climate model level and the climate simulation level as for CMIP5.

3.5 Additional Requirements for CMIP6 Archive Centers / DataCite DOI Data Centers

- Long-term archived data
- Long-term and permanent data access
- Detailed documentation for interdisciplinary data use (enrichment of ESGF index metadata by CIM documents, quality information, annotations if available)
- IPCC-DDC: stability of data (data is complete and persistent and suitable for use in a peer-review data journal); Core data definition required for timely integration of the data in the IPCC-DDC including DataCite DOI data citations: DECK+ experiments (DECK+ScenarioMIP+…) and important variables

3.6 Additional Requirements on Discovery

Data is one entity in the research environment of CMIP6. Therefore the different entities should be connected and related to each other in a transparent and understandable way.

- ***DOI Data Discovery*** (DRS, PIDs on related entities):
  The meta-data record submitted to the DataCite registry should contain enough information to enable effective data discovery through the DataCite metadata search. Institutions and key individuals should be properly identified using the designated mechanisms (e.g. ORCID, FundRef, etc.). The relation of the AR6 data subset to the CMIP6 data should be clearly specified in the DataCite metadata. Other related information entities with PIDs can be treated, accordingly. Additional vocabularies for CMIP6 should be integrated to enable a more precise mapping of Data Reference Syntax terms into DOI metadata. DRS terms will be part of the DataCite metadata, ideally with a reference to the CV catalog of CMIP6 DRS terms.
- ***ESGF Data Discovery*** (DRS, PIDs on related entities):
  In the ESGF portal any additional information on the data improves the reusability of ESGF data. At the moment CIM documents are linked in the portals. With the change to the CoG portal, additional repositories plan to loosely link to their information, e.g. citation, errata. This information should be available for ESGF replication or search tools. A connection of these entities by PIDs would be desirable, but at least DRS_ids are required for the identification of related documents.

Table 2: Overview over requirements for CMIP6/AR6 data citations and its targeted solutions; data paper publication/citation are an option for modeling centers to add value to their data citation

	Criteria	**CMIP6 Data Citation (in ESGF with HDL Identifier)**	**IPCC-AR6 Data Citation (in WDCC-LTA with DOI)**	**Citation of Data Paper (in data journal with DOI)**
3.1	Credit and Attribution (F11)	for creators and contributors	for creators and contributors	for creators and contributors (ESSD indexed by Thomson Reuters in Web of Science)

3.2	Unique Identification, Access and Persistence (F11)	with HDL Identifier; dynamic data (content changes);	with DOI; stable and complete data	with DOI on paper and DOIs (or other PIDs) on referenced data
3.3	Specificity and Verifiability (F11)	requires versioning and version support; few information on data usage	stable data with permanent access and detailed usage information	stable and complete data required with PID access
3.4	Data Aggregation	model and simulation	model and simulation	model
3.5.1	Long-term Data Availability	Sufficient dataset lifespans in fixed data collections (persistence of metadata, PID, and provenance information)	data unchanged and permanently accessible on the long-term	data stable, i.e. unchanged
3.5.2	Data Stability*	datasets are finalized but not persistent; dataset collections are not complete and have changing content	data is finalized, long-term archived and curated	finalized and persistent
3.6.1	Data Quality / Documentation	quality assurance of conformance; linked documentations available	informal scientific review by community; add. technical quality assurance; comprehensive documentation	peer-review; article as additional detailed documentation; documentations of data repositories.
3.6.2	Relations to other entities	In ESGF portal links to additional data can be implemented, e.g. CIM, citation, errata, quality; additional repositories should provide references to ESGF data	Additional data can be stored or referenced (if PID is provided); DRS integration in DataCite metadata; DataCite metadata integrated in multiple catalogs, e.g. of publishers.	data journals are indexed by publisher services

**: Data stability (completeness and persistence) is not explicitly covered in Force 11 but required by publishers of scientific journals. For observational data completeness a certain and documented date is sufficient. Data might be added later. In CMIP the datasets are revised, partly (or in certain cases even completely) deleted.*

References:

Earth System Science Data (ESSD): Repository Criteria, http://www.earth-system-science-data.net/general_information/repository_criteria.html.

Geoscience Data Journal: Guidelines for Repositories, http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060/homepage/guidelines_for_repositories.htm.

Nature Scientific Data Journal: Data Policies, http://www.nature.com/sdata/data-policies.

4. Implementation / Workflow

Citation information provided by the modeling centers will be used for both citations: dynamic CMIP6 data and stable AR6 data subsets. The user view on the two data citations for the two data collections should be as transparent and uniform as possible. Thus, citation recommendations and policies should appear on the landing pages for both citations. However, data citation principles require the clear separation of the cited data from the data related to the cited data.

4.1 Workflow

Selected data node managers assigned by CMIP Data Node Operations Team (CDNOT; selected data nodes are called ***'QA data nodes'*** in this document) are responsible for quality assurance of citation information content and its compliance to data citation policies. The citation repository provides a GUI for citation information update and an API for citation information access.

Workflow steps:

Steps 1 -7 represent the early citation workflow for dynamic CMIP6 data, while steps 8 – 9 belong to long-term archiving of CMIP6/IPCC-AR6 reference data including DataCite data publication.

1. Initial (default) citation information collected by WIP from modeling centers on model granularity: scientific contact person provided within metadata
2. Initial citation checked and inserted in citation repository by WDCC for model and simulation granularities
3. Citation repository has API in place to provide information for ESGF and other CMIP6 repositories, e.g. CIM
4. Initial ESGF data publication: check on availability of citation information performed prior to data publication
5. 'QA data nodes' check data citation content
6. Citation repository GUI for changes of citation contents in place
7. Changes of citation information:
    a. Modeling center inserts content in citation repository via GUI (support of WDCC for significant content changes)
    b. Modeling center sends revised data to its responsible data node and 'QA data node' checks citation content and
    c. citation changes are immediately available via API for ESGF and other CMIP6 repositories
8. Replication of Core data for long-term archival (LTA) and IPCC-DDC AR6 including related content out of external repositories (as available): CIM, Errata, QC,…
9. LTA of data and metadata
10. Content on landing pages extended by replicated metadata
11. Scientific contact person of modeling center is contacted for final author approval
12. Citation information is fixed for CMIP6 and AR6
13. DataCite DOI publication and transfer into IPCC-DDC AR6

4.2 Citation Policies

4.2.1 Data Citation Content Policies

Citation information is partly collected from the modeling centers like creators and titles, other parts are restricted by policies and set by the publisher, like publisher and publication year (see Table 3).

Table 3: Citation information content

Citation Content	**mandatory/ optional**	**CMIP6 citation**	**AR6 citation**	**policy**	**responsible parties**
Creator list	mandatory	list of institutions or persons	list of institutions or persons	CMIP6: can change[3], AR6: fixed; Identical for CMIP6/AR6	modeling centers (QA: CMIP6: QA data nodes[4] AR6: WDCC)
scientific contact	mandatory	member of creator list or other person	member of creator list or other person	Identical for CMIP6/AR6	modeling centers (QA: CMIP6: QA data nodes[4] AR6: WDCC)
Publication year	mandatory	Year of initial ESGF publication	Year of initial ESGF publication	Identical for CMIP6/AR6	CMIP6: QA data nodes[4] AR6: WDCC
Title	mandatory	defined by modeling centers	defined by modeling centers	Identical for CMIP6/AR6	Modeling centers and CMIP6: QA data nodes[4] AR6: WDCC
Publisher	mandatory	Earth System Grid Federation	World Data Center Climate at DKRZ	Different for CMIP6/AR6	WDCC
PID	mandatory	HDL Identifier	DataCite DOI	same PID suffix, same content for landing pages	WDCC
Contributor lists	optional	list of institutions or persons	list of institutions or persons	CMIP6: can change, AR6: fixed Identical for CMIP6/AR6	modeling centers (QA: CMIP6: QA data nodes[4] AR6: WDCC)
Person Information (creator/ contributor)	mandatory/ optional	mandatory: name, email, institute; optional: ORCID, etc.	mandatory: name, email, institute; optional: ORCID, etc.	Identical for CMIP6/AR6	modeling centers (QA: CMIP6: QA data nodes[4] AR6: WDCC)
Institute Information (creator/ contributor)	mandatory/ optional	mandatory: name, acronym, URL; optional: address, etc.	mandatory: name, acronym, URL; optional: address, etc.	Identical for CMIP6/AR6	modeling centers (QA: CMIP6: QA data nodes[4] AR6: WDCC)
Contributor type (e.g. Funder, Researcher)	mandatory for contributor lists			Identical for CMIP6/AR6	modeling centers (QA: CMIP6: QA data nodes[4] AR6: WDCC)
Project Description (abstract) on landing pages	mandatory as initial landing page content	mandatory (more information added as available)	mandatory (more information added as available)	Identical for CMIP6/AR6	WDCC (project description provided by WIP)

[3] After final author approval (workflow step 11) the creator list gets fixed.

[4] These data nodes assigned by CDNOT should be responsible for the quality assurance of citation content (e.g. meaningfulness of titles, correctness of person names, etc.) and the compliance to data citation policies on their share of the overall CMIP6 data published in ESGF.

4.2.2 Data Citation Usage Recommendations

CMIP6 data used in publications or presentations by other scientists should be cited in general. As the ***granularity for data citations*** has to be suitable to balance the number of data entries in the reference list with paper entries, two aggregations are offered:

- ***Citation of model data:*** intercomparison of multiple models and multiple experiments,
- ***Citation of simulation data:*** analysis of few experiments from one or multiple models.

Depending on the needed granularity for citation the source and time of data download, either the CMIP6 data (downloaded from ESGF) or the IPCC-DDC AR6 data (downloaded from WDCC) two closely linked data citations are recommended with PIDs pointing to landing pages of (nearly) identical content (cf.).

1. Citation of Dynamic CMIP6 Data:
   data creator list (publication year): Title. Version. Publisher. HDL Identifier.
2. Citation of Stable IPCC-AR6 Data:
   data creator list (publication year): Title. Publisher. DOI.

The ***recommendation for usage of CMIP6/AR6 data citations*** should be:

- ***CMIP6 Data Citation:*** The use of dynamical CMIP6 data is recommended during CMIP6 project phase (IPCC-DDC data not yet available). The CMIP6 data citation should be additionally recommended in case the used data version differs from the AR6 data version or is unknown.
- ***AR6 Data Citation:*** The use of IPCC-DDC AR6 data is recommended for interdisciplinary reuse on the long-term.

4.2.3 Landing Page Content

The same layout and content is provided on both landing pages. This content is also shown for the ESGF portal users accessing the 'Data Citation' link. As more metadata becomes available, the content is added. As initial content, information on the project is provided with links to further

Landing Page Sections	Content CMIP6/AR6
DRS name	Identical
Citation information Data Access link Related Data: Citation information Data Access link for related data	Order of citations/ access links depending on PID
Citation Policy/Usage Information Further information (at least on project)	Identical

information.
The order of the citation information / data access links changes depending on the accessed PID / the data citation. The citation policy and usage section explains their relation and recommends the data citation for the different users (section 4.2.2).

4.3 CV for DRS Terms

A central CV for all DRS terms is required to grant stable and uniform use of those names during data creation (by CMOR2), data distribution (in ESGF) and in external infrastructure components providing additional information on the data (CIM documents, quality information/results, citation regulation, etc.). Aliases for directory names and search facet names are needed.
The currently used loosely connection of CIM documents to ESGF and planned connection of QC and citation information is based on DRS terms to identify related contents. Therefore the uniform use of DRS terms by all technical infrastructure components is crucial.

For use in the DataCite metadata a ***persistent URI*** is needed as reference for the DRS terms CV.

4.4 ESGF Aspects for CMIP6 Citations

A data user resolving a PID of a data citation is redirected to the PID landing page, where access to the data in the ESGF portal is provided. With the version[5] information in the citation string, the user needs to be able to restrict the results in the ESGF portal from all dataset versions to the cited data (model or experiment data), which are the latest dataset versions for the past date. A possibility to compare the data of the cited (past) version to the latest (current) version is desirable.

Thus the ESGF needs to support the versions in the search functionality and in the ESGF CoG portal for the ***version use cases***:

- "Get latest dataset versions for a search query for a specific date in the past"
- "Show differences of the latest dataset versions for a search query for a specific date in the past to the current date"

4.5 Replication and LTA Aspects

For the long-term archival the replication of data and metadata in latest versions for a specified date is required. The replication needs to support the ***replication use cases***:

- "Get latest dataset versions for a specific date for a specified subset of the CMIP6 data (e.g. a model)"
    - Option: "Get additionally related metadata"
    - Option: "Use predefined data specifications to replicate a data subset (Core data subset support)"

The LTA investigates how to integrate the ***additional metadata*** into the existing metadata scheme and whether an external metadata service could be referenced, which requires persistent access using DRS terms. At least spot checking to validate the relation of the external metadata to the data is performed prior to the start of the DataCite DOI publication process. If possible cross-checks are added to the TQA (technical quality assurance) for automated checking.
A question to answer is whether there will be ***CMIP6 Archive Centers*** storing replica of the most

[5] Dataset version strings should include the date of creation or ESGF publication.

important ***CMIP6 data (Core)*** and the integration of ESMVal variables into the Core. A starting point for definition of a Core data subset could be the IPCC AR4 and the IPCC AR5 WG1 subsets.

4.6 DataCite DOI Aspects

The identification of CMIP6 data in the DataCite catalog and catalogs which harvest the DataCite metadata should be possible. That requires that the DRS terms are published in the DataCite metadata for a citation entity. The 'subject' section in the DataCite XML will be used, ideally with a reference to the DRS CV using the attributes 'schemeURI' and 'subjectScheme'.

4.7 Citation information access (API)

Citation information will be made accessible via an API using DRS terms with display and data options for user and machine access (formats to be decided, e.g. XML, JSON; Figure 2). The information is structured as DataCite metadata (allowing additionally the use of HDL Identifiers).

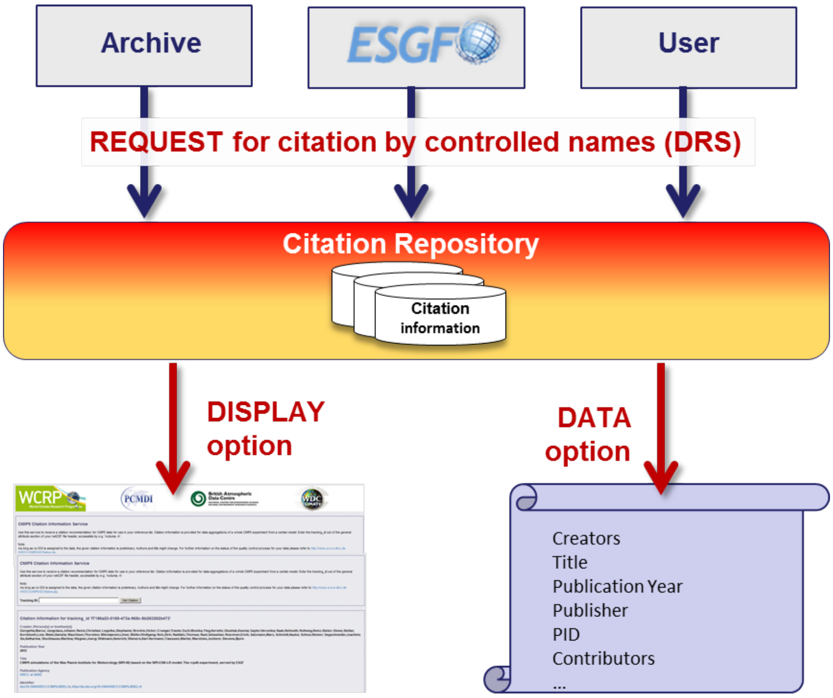


Figure 2: Citation API for citation information access.

The display option will redirect to the landing page. A link 'Data Citation' is to be added on the ESGF portal on dataset level using the API with display option. This integration of citation information is coordinated within the ESGF-QCWT (ESGF Quality Control Working Team).
The target page of the furtherInfoUrl[6] will harvest citation information and display it along with other additional information.

4.8 Timeline

Q2/2015: WIP agreement on data citation concept

[6] furtherInfoUrl is a global attribute in the file header providing a persistent reference to further information (see WIP paper on file names and global attributes for more information).

Q3/2015: Finalize Citation Repository structure;
Modeling centers asked for contacts and initial citation information on model data

Q4/2015: Finalize Citation Service development: prototypes for GUI, API, and landing page; Integration of Citation API into ESGF CoG portal

Q1/2016: Testing of Citation Services

Q2/2016: Operable Citation Services;
Adaptation LTA and DataCite DOI processes to CMIP6 requirements

Q4/2016: Testing of LTA and DataCite DOI processes finalized

Q1/2017: Operable long-term archival procedure and DataCite DOI process

Reference:

DataCite Metadata Schema 3.1, doi:10.5438/0010.

5. Connections to other WIP white papers and groups

- **CMIP Data Node Operations Team (CDNOT):**
  - Integration and supervision of early CMIP6 data citations in the ESGF publication process incl. QA aspects and responsible persons
  - Integration of different information components (CIM, quality, errata, citation) into ESGF data infrastructure incl. accessibility by DRS terms (or PID services)
  - Integration of related metadata (CIM, quality, errata, citation) into ESGF replication software
  - Version support in ESGF is crucial for the citation of dynamic CMIP6 data (see use cases in sections 4.4 and 4.5)
  - Ensure an operable overall technical infrastructure with reliable services
  - CMIP6 Archive Centers storing Core data subsets as replica: Core data have to be defined preferably before CMIP6 starts or at the very beginning of the process. There are relations to the ESMVal variables as well as IPCC AR5 WG1 data subset (experiment, variables).

- **WIP white papers:**
  - *Quality/Errata/CIM:* A certain data and metadata quality is defined there as requirement for long-term archival and IPCC-DDC integration. The archival of CIM model and simulation documents as well as quality issues is desirable.
  - *Versioning:* Strict data versioning and the persistence of ESGF metadata as well as version support in the ESGF portals are requirements for data citations.
  - *Replication:* Replication of data and related metadata from external repositories is the requirement for the long-term archival of the data and integration into IPCC-DDC.
  - *Licensing and Access control:* The integration of data into the IPCC DDC AR6 requires an open access data license.
  - *Data Reference Vocabulary*: There are dependencies on definition and access to CVs (CVs for DRS terms are of critical importance) as well as on the definition of an important subset of variables for CMIP6 Archive Centers and its accessibility as machine-readable list.

  - *File Names and Global Attributes:* The citation information should be integrated in the target page of the furtherInfoUrl, which is a global attribute in the file header.
  - *Discussion on data volumes:* high dependency for Core data definition of IPCC-DDC so that the data citation can meet the IPCC-AR6 timeline
  - *On Specification and documentation of data, models and experiments*: access to additional information on data (CIM documents)
  - *Fine-granular PIDs*: PIDs on ESGF files and datasets should be integrated into the metadata of the long-term archived data if available.

- **WGCM:**
  - *An ESSD special issue on data in a CMIP6 Archive Center or on IPCC AR6 data* (to be discussed) would require a co-ordination of data long-term archival/data citation and IPCC AR6 workflows and schedules required.