DATA SCIENCE JOURNAL

Stockhause, M and Lautenschlager, M 2017 CMIP6 Data Citation of Evolving Data. *Data Science Journal,* 16: 30, pp. 1–13, DOI: https://doi.org/10.5334/dsj-2017-030

RESEARCH PAPER

CMIP6 Data Citation of Evolving Data

Martina Stockhause and Michael Lautenschlager

WDC Climate at German Climate Computing Center (DKRZ), Hamburg, DE

Corresponding author: Martina Stockhause (stockhause@dkrz.de)

Data citations have become widely accepted. Technical infrastructures as well as principles and recommendations for data citation are in place but best practices or guidelines for their implementation are not yet available. On the other hand, the scientific climate community requests early citations on evolving data for credit, e.g. for CMIP6 (Coupled Model Intercomparison Project Phase 6). The data citation concept for CMIP6 is presented. The main challenges lie in limited resources, a strict project timeline and the dependency on changes of the data dissemination infrastructure ESGF (Earth System Grid Federation) to meet the data citation requirements. Therefore a pragmatic, flexible and extendible approach for the CMIP6 data citation service was developed, consisting of a citation for the full evolving data superset and a data cart approach for citing the concrete used data subset. This two citation approach can be implemented according to the RDA recommendations for evolving data. Because of resource constraints and missing project policies, the implementation of the second part of the citation concept is postponed to CMIP7.

Keywords: CMIP6; IPCC-DDC; climate data; data publishing; evolving data; earth system modeling

1. Status of CMIP Data Citations and Evolving Data Citations

1.1 Citations of Static and Evolving Data

Over the last decade a rapid standardization process for data citation in parallel with data publication practices took place. Starting 2004 as a German project, the foundation of DataCite on 1 December 2009 transformed this national effort into a successful international development (Brase et al., 2015). The FORCE11 Data Citation Synthesis Group formulated the 'Joint Declaration of Data Citation Principles' as general guidance on purpose, function and attributes of data citations (FORCE11, 2014; Altman et al., 2015; **Table 1**). The FORCE11 FAIR Data Publishing Group built on these the FAIR (Findable, Accessible, Interoperable, Re-usable) principles for data sharing (FORCE11, 2017). Additional standardization initiatives investigate the data publishing workflows and develop reference models, like the RDA/WDS Workflows Working Group (Austin et al., 2016). Others focus on data review and its integration in the scholarly publishing processing (Mayernik et al., 2015).

For these approaches static data is assumed, which is long-term archived after the termination of the scientific project. However science is innovative by nature and thus research data as a product of science is dynamic, especially high volume data released early in the project, like in the Coupled Model Intercomparison Project (CMIP, 2017). Klump et al. (2016) suggest to characterize the kind of dynamic data changes as growing, evolving and fragmented datasets, where data is appended for growing datasets, data is revised and new versions are appended for evolving datasets, and data is partly replaced for fragmented datasets. Hereafter, only growing and evolving datasets will be discussed using the more general term evolving data. The RDA Data Citation Working Group (WGDC) agreed on a set of 14 RDA-endorsed recommendations for citations of evolving data (Rauber et al., 2015; Rauber et al., 2016; **Table 1**), which has become part of the DCC (Digital Curation Centre) guideline on data citation (Ball and Duke, 2015). The concept of a Persistent Identifier (PID) generation on demand or when the data gets cited, underlie this approach.

Table 1: General Force11 data citation principles (left) and recommendations of the RDA WGDC on the citation of evolving data (right).

Force 11 – Data Citation Principles	RDA-endorsed recommendations on the citation of evolving data (RDA WGDC)	
1. Importance	R1 – Data Versioning	A. Preparing the Data and the Query Store
2. Credit and Attribution	R2 – Timestamping	
3. Evidence	R3 – Query Store	
4. Unique Identification	R4 – Query Uniqueness	B. Persistently Identify Specific Data sets
5. Access	R5 – Stable Sorting	
6. Persistence	R6 – Result Set Verification	
7. Specificity and Verifiability	R7 – Query Timestamping	
8. Interoperability and Flexibility	R8 – Query PID	
	R9 – Store Query	
	R10 – Citation Text	
	R11 – Landing Page	C. Upon Request of a PID
	R12 – Machine Actionability	
	R13 – Technology Migration	D. Upon Modifications to the Data Infrastructure
	R14 – Migration Verification	

ARGO (2017) is one among several early implementers of these recommendations using DataCite DOIs with standard #-fragments (see section 5.8 of IDF DOI Handbook, 2017). These fragment identifiers are technically supported by DataCite but are currently lacking guidance on its usage. ANDS (2017) recommend the usage of @-fragment identifiers for growing and evolving data. The use of fragment identifiers is a controversial issue, because it partly contradicts the idea to separate changing URLs from persistent IDs by introducing an addition to the PID, which is interpreted server-side. Another common approach to handle evolving data is the use of full or partial snapshots of the data with individual registered PIDs (Ball and Duke, 2015). Snapshots are suitable for low volume data and defined rare data changes, like evolving observation data with yearly snapshots. For the high data volumes of CMIP the use of snapshots is not feasible.

The present discussions are driven by scholarly publishers and infrastructure providers, like the European Persistent Identifier Consortium (EPIC). They concentrate more on Force11's data citation principles for verification and unique identification than on the principle for credit and attribution. More important than citing one's own data as verification for the findings of an article is to cite the data of other scientists, on which the findings of the article and the derived datasets are based. Lawrence et al. (2011) suggest the use of two identifiers for data citation within a single reference by introducing an optional featureID to specify a data subset in addition to the PID of the cited data. The use of two citations, one for the data superset giving credit to the creators and a second for the data subset identifying the used data, is implicitly included in the RDA-endorsed recommendations for citations of evolving data. Fenner (2016) recently suggested adding an explicit recommendation R15 "PID on whole dataset" with a metadata reference to the related PID on the full dataset in the metadata of the data subset.

In this paper will refer to the two data citations as ***Data Superset Citation/PID*** and ***Data Subset Citation/PID*** to distinguish references on the evolving full dataset or data superset (data referenced > data used) from references on used data subsets (data referenced = data used). The data supersets are used within reference lists of scholarly publications and enable to give credit to the data creators and to provide statistics on data usage in literature. The data subset is used for the identification of the part of the data underlying an article.

1.2 Data Citation Situation in CMIP5

The data publishing situation in Earth System Modeling has aspects of the "Big Iron" metaphor (in data characteristics, cultural context, standard emphasis) and the "Data Publication" metaphor (in data quality/thoroughness of metadata annotations, long tradition of data publishing) proposed by Parsons and Fox (2013). Arguments raised in current data publishing discussions are based on both metaphors. Data management and identification issues are technically discussed along the "Big Iron" metaphor, e.g. by EPIC. The data citation in scholarly literature as use case and arguments like credit for data creators

are exchanged along the "Data Publication" metaphor between data publishers and data registries like DataCite. A data citation approach for Earth System Modeling has to incorporate and give answers within both metaphors.

Climate research is characterized by high volume, globally distributed data and multiple contributing researchers and research institutions. E.g. about 30 institutions provided data with 60 models for the last Coupled Model Intercomparison Project Phase 5 (CMIP5) to build a data base of ca. 1.6 PBytes, which underlies the 5th Assessment Report on the world's climate of the Intergovernmental Panel on Climate Change (IPCC AR5) and is part of the IPCC Data Distribution Centre's (IPCC-DDC, 2017) Reference Data Archive. The IPCC-DDC Reference Data hosted by the WDC Climate at DKRZ facilitates the long-term interdisciplinary usage of the data beyond the individual CMIP project phase and the individual IPCC assessment cycle. The IPCC-DDC AR5 data collections are quality assessed and citable by DataCite DOIs (Stockhause et al., 2012). The citation granularity is a simulation consisting of several ensemble runs for a scientific experiment. The number of individual datasets per DOI was in the order of 100 to 1000.

This assignment of data citations for long-term archived data has become a standard service of established data archives such as the WDS (World Data System) certified archives. After the end of a research project, the finalized (static) data is transferred into a long-term archive and minted a DataCite DOI making the data citable. These DOIs are data superset PIDs according to the definition above.

However in case of CMIP5 data, the first datasets were disseminated years before the end of the project. Therefore data was downloaded, analyzed and used in scholarly publications without citing the underlying data. This was also true for the three parts of the IPCC AR5.

2. Data Citation Concept for CMIP6 and IPCC-DDC AR6

The experience for CMIP5 showed that data citations need to be provided not only for the long-term archived IPCC-DDC data but also for the still evolving CMIP6 project data. The challenge for the data citation service is to close this gap in cooperation with the data infrastructure provider, the Earth System Grid Federation (ESGF, 2017), and in the timeframe given by the CMIP6 project and IPCC. Currently, the fixed date for the submission deadline for 6th Assessment Report (AR6) of Working Group I is March 2020. Therefore the analysis of CMIP6 data by the IPCC authors will start about two years earlier in early 2018 and the first CMIP6 datasets are expected to be ESGF published around mid of 2017.

2.1 Developments in the CMIP Project

CMIP is coordinated by the Working Group on Coupled Modeling (WGCM, 2017) of the World Climate Research Programme (WCRP). For CMIP6 the structure of CMIP was changed from separate phases to a more continuous approach. Thus, the 21 included individual Model Intercomparison Projects (MIPs) of CMIP6 can start at different times and can utilize existing evaluation simulations. An overview over CMIP6 is given by Eyring et al. (2016). Estimations for the overall CMIP6 data volume range from 15 to 30 PBytes of binary data in the self-describing community format NetCDF.

The WGCM Infrastructure Panel (WIP, 2017) was formed to coordinate the development of the federated data and metadata infrastructure and formulate policies for data node managers. The WGCM CMIP6 (2017) requested an additional data citation possibility for CMIP6 data in order to give credit to the data creators, their funders and other contributors.

2.2 Data Structure and Data Infrastructure Developments

The data infrastructure ESGF was developed during CMIP5 data dissemination superseding the ESG (Earth System Grid). With the major software overhaul in the second half of 2015 the ESGF the main development was finalized. The ESGF architecture is based on a P2P (Peer-to-Peer) system of autonomous and worldwide distributed nodes providing different services. CMIP data is stored over multiple data nodes. Metadata is exchanged among the index nodes, enabling data discovery over the complete CMIP data hosted at multiple data nodes of the federation.

Citation information is not available within the ESGF. This information is stored in a separate central repository. For the integration of citation and other ancillary metadata information, an ESGF registration functionality was developed in collaboration with several ESGF Working Teams.

Another ESGF development crucial for the citation service is the version support. Each dataset published in the ESGF gets versioned with the publication date. Within CMIP5 dataset versions in ESGF were regarded as attributes. The publication of a new dataset version overwrote the metadata and an unpublication deleted all information. ESGF metadata is changed into persistent metadata enabling the discovery of unpublished datasets, which are clearly flagged in the search result list.

As the ESGF infrastructure disseminates data of different projects, the infrastructure development is accompanied by CMIP6 project agreements. Data will be replicated within the federation among the tier1 and tier2 data nodes ensuring at least a second or even a third copy of the CMIP6 data. Negotiations about the concrete replication strategy are ongoing.

Data within CMIP has to comply with naming conventions, which are available as set of controlled vocabularies.[1] These names are components of the data reference syntax. They are used as directory structure for data storage on disk as well as for the unique identification of datasets including versions (dataset ID). Datasets consist of multiple NetCDF files and have the structure (Taylor et al., 2017):

<mip_era>/<activity_id>/<institution_id>/<source_id>/<experiment_id>/<member_id>/<table_id>/<variable_id>/<grid_label>/<version>

e.g.: CMIP6/CMIP/NOAA-GFDL/GFDL-CM2-1/1pctCO2/r1i1p1f1/Amon/tas/gn/v20150322.

The components of this structure are ESGF search facets and allow for the identification of datasets and files by content, such as institute, model, experiment, run, frequency, etc. A query for such an individual dataset requires the specification of the values for all 10 search attributes. Alternatively, the dataset ID of the dataset can be used in the search query.

2.3 CMIP6 Data Citation Concept

CMIP data is evolving. The data is published in the ESGF when it becomes available. Thus the CMIP repository is building up in the first phase, e.g. data of model runs or new post-processed variables are added. In the second phase individual variables across multiple runs are corrected and ESGF-published under new versions. This stabilization process of CMIP data takes several years.

Data citations are designated on data aggregations belonging to a model contribution to a CMIP6 MIP and on data belonging to an experiment contributed by a specific model:

model citation: <mip_era>/<activity_id>/<institution_id>/<source_id>
experiment citation: <mip_era>/<activity_id>/<institution_id>/<source_id>/<experiment_id>.

The number of datasets included will be larger than in CMIP5 reaching from the order of 1000s to 10000 datasets. The two citation granularities are offered to keep the balance between data and literature citations in reference lists of scholarly publications for different research studies. Typically, few variables out of the large data collections are analyzed across several runs or in intercomparison studies across several runs and several models. Thus data subsets cited in scholarly literature are still large collections of individual datasets and they typically include datasets from several data supersets (citation entities) in specific versions.

In comparison with the long-term archived IPCC-DDC data, CMIP data has a lower curation standard and is less well-documented. Single variables within the data aggregation of a citation entity might be deleted. And CMIP6 data is stored at several data repositories worldwide. The lower curation standard of the CMIP6 data needs to be visible to data users. Therefore it was agreed to mint DataCite DOIs for CMIP6 data on behalf of the data publisher ESGF. Another DOI prefix '10.22033' is used to distinguish both within DataCite and to reserve the possibility to transfer ESGF data citations to a future ESGF member of DataCite.

The version is part of the citation recommendation for the evolving CMIP6 data, which is based on the ESGF publication date:

*Authors/Data Creators (publication year): Title. Version **YYYYMMDD**. Publisher. DOI.*

This version is the latest version of any static dataset included in the cited data collection (model or experiment). If this latest dataset version is unknown, the download date can be used instead. The recommended citation for CMIP6 data is conformant with the ESIP recommendations (ESIP Stewardship Committee, 2012). The ESGF portal supports filtering for versions between certain given dates.

A CMIP6 data subset at a certain snapshot date will be transferred into the IPCC-DDC to build the Reference Data Archive for the 6th Assessment Report (AR6). The content of the CMIP6 data subset will be defined together with the IPCC Working Groups. These data will be issued DataCite DOIs by *WDC Climate* on the same granularities as for the evolving CMIP6 data. Relations between CMIP6 and IPCC-DDC AR6 data citations will be published as part of the citation metadata (**Figure 1**). Both citation DOIs function as data superset PIDs. The terms of use for data citation recommend the use of IPCC-DDC AR6 citations if available.

[1] Controlled Vocabularies for use in CMIP6 are accessible at: https://github.com/WCRP-CMIP/CMIP6_CVs.

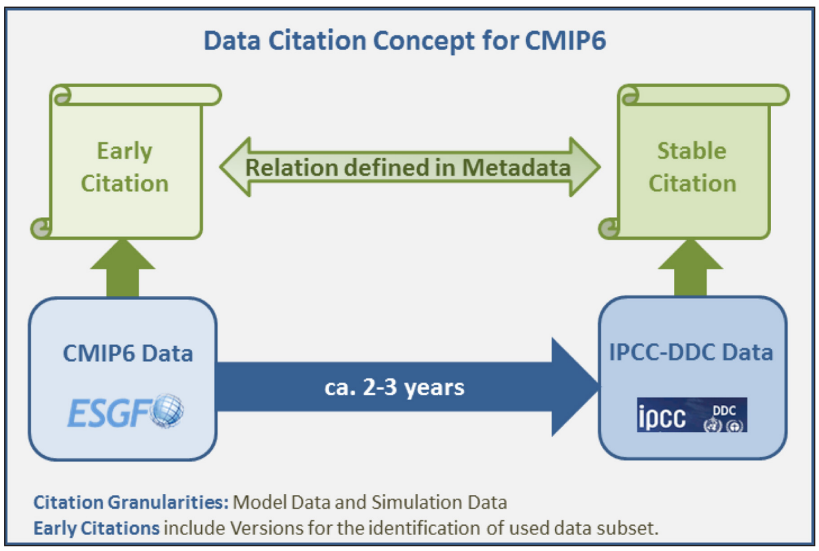

Figure 1: Relation between CMIP6 early data citations and the IPCC-DDC AR6 Reference Data Archive in terms of data citations. The IPCC-DDC AR6 data is a snapshot as well as a subset of the CMIP6 data.

2.4 CMIP6 Data Citation Implementation

The CMIP6 data citation concept (Stockhause et al., 2015; CMIP6 Citation Service, 2017) has to meet the three main use cases for a data citation service: provide citation information, discover citation information, and resolve data citations (**Figure 2**). It was approved by the WIP as part of the CMIP6 infrastructure.

The citation information is stored centrally in a database schema in the Oracle database of the WDC Climate (WDCC). The scheme utilizes the information for persons and institutes of WDCC's long-term archive. Citation details are entered by data creators via a Graphical User Interface (GUI) based on Oracle APEX (Application Express). Information on creators, contributors, title but also on references to related scholarly literature is collected. The provision of ORCID researcher IDs (Orcid, 2017) for persons is recommended. The citation service provides all information related to a data citation on the landing page (see example in **Figure 3**).

The displayed data access links on the landing page point into the Earth System Grid Federation portals and use ESGF search queries to filter the available data for the datasets belonging to the data citation entry, i.e. experiment or model data. It is currently under discussion whether to use fragment identifiers by adding the version to the end of the DOI. This version information has to be interpreted by the Citation Service to extend the data access link displayed on the landing page. The advantage for a user resolving the data superset DOI including a fragment identifier extension is that the dataset list in the ESGF is already filtered for versions which were available at the given date or for the given cited dataset versions.

Citation information is exchanged as DataCite metadata (DataCite, 2016) in XML and JSON formats. The different formats HTML, XML and JSON are accessible under the same URL using content negotiation. The default format is HTML or the landing page. JSON format is used by the ESGF to display core citation information in the ESGF portal for data users (**Figure 4**). During data publication a citation link is registered in the ESGF index, which is underlying the "Show citation" link for every dataset. Citation information is dynamically loaded upon user request by sending a request to the citation API and rendering the core citation information of the returned JSON file for display. A link to the landing page with the complete citation information is part of the displayed data citation information. The registration of the citation link in the ESGF index enables the exchange of citation information among ESGF index nodes as well as the access of this ancillary metadata by external services like the long-term archival by the IPCC-DDC.

A second user possibility to access citation information is provided via the "furtherInfoUrl" link in the NetCDF file headers. Its landing page hosted by Earth System Documentation (ES-DOC, 2017) is intended to display all kinds of available ancillary metadata including ES-DOC information on model, simulation etc. and data citation details.

2.5 Integration of CMIP6 Data Citations into the Scholarly Environment

Apart from the integration of the CMIP6 Data Citations into the CMIP project, data citations are also part of the scholarly environment. The Scholix framework (Scholarly Link eXchange; Scholix, 2017) develops visions and guidelines for linking research data and literature to increase interoperability. The RDA/WDS Scholarly

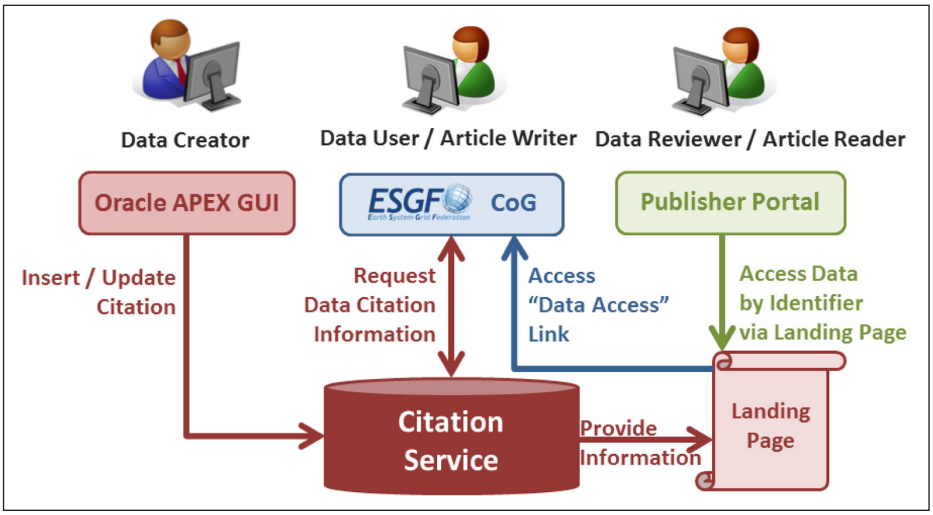

Figure 2: CMIP6 Data Citation Concept with the three main use cases to meet: provide citation information, discover citation information, and resolve a data citation (red: CMIP6 citation services, blue: ESGF services, green: services of scholarly publishers).

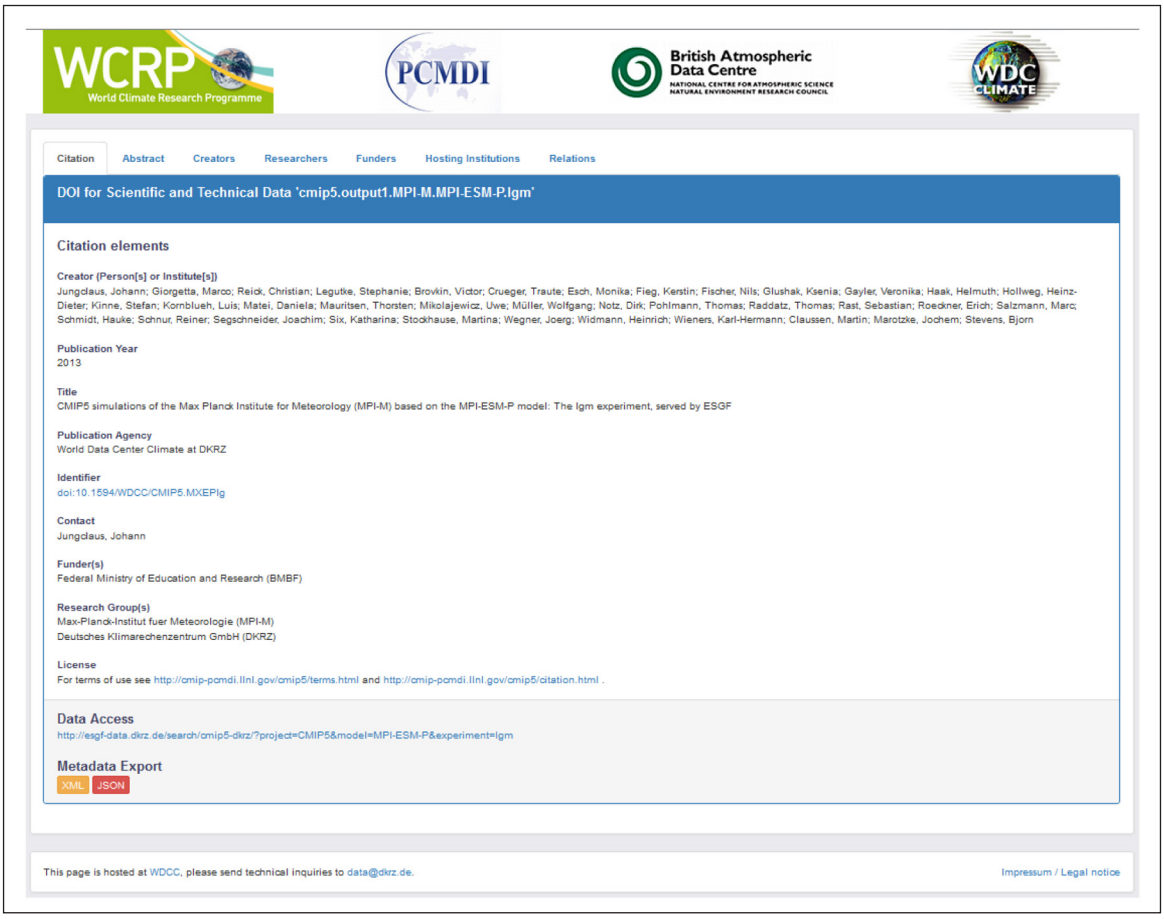

Figure 3: Landing page example for CMIP6 early citation based on CMIP5 data as proof of concept.

Link Exchange Working Group coordinates the implementation of these Scholix guidelines. The CMIP6 and IPCC-DDC AR6 data citations will automatically profit from these improved data-data, data-literature and data-researcher linkages via DataCite as hub. The DataCite DOI metadata of AR6 data citations will include relations to CMIP6 data citations, to known scholarly publications and researcher profiles via ORCIDs to enable interlinking. The IPCC-DDC on the other hand is interested in harvesting information about the usage of its data in scholarly literature and in providing statistics on data usage for the CMIP6 data creators.

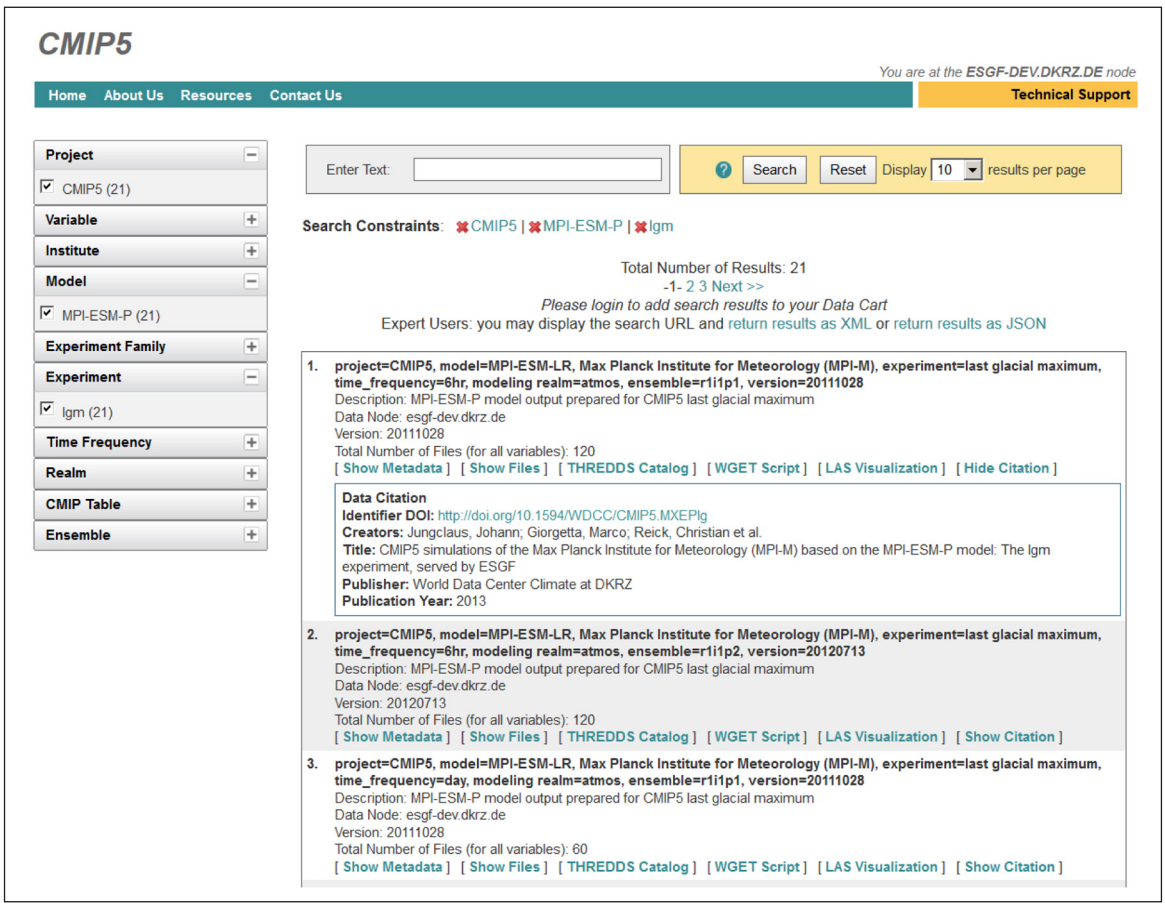

Figure 4: ESGF portal with display of data citation information for a CMIP5 data example as proof of concept.

For direct metadata harvesting, e.g. by OpenAire (2017), CMIP6 and AR6 citations will be provided as set of DataCite metadata XMLs on WDCC's OAI Server.

3. Discussion of CMIP6 Data Citation Approach

Comparing the CMIP6 citation approach for the data superset DOIs described in section 2 with FORCE11's Data Citation Principles (**Table 1**), the principles are satisfied but only for humans not for machines. As these data superset DOIs include more data than used within the scholarly publication, identification and access of the individual datasets of the data collection used in the publication (principles 5 and 7) require reading the main article. Principles 1 to 4 connected to importance, credit and evidence are fulfilled for CMIP6 citations as the importance and requirement was pointed out by the WGCM CMIP6. The unique identification and interoperability/flexibility (principles 4 and 8) are given by the use of DataCite DOIs. As CMIP6 data is evolving, the content will change and some individual datasets of the collection might no longer be available. However the collection itself, the metadata and the landing page persist, complying with principle 6 on persistence.

Since the persistent identification of the used data subset is not available for the described first implementation step of the CMIP6 citation concept, the RDA-endorsed recommendation on the citation of evolving data are not applicable. Nevertheless a comparison against the recommendations R1–R3, concerning the preparation of the data and query store, provides a useful estimate of the effort needed for complying. ESGF stores versions for static datasets in its metadata and provides unique dataset IDs using a naming convention. Versions are the dates of the ESGF publication of a dataset. Cited data subsets consist of multiple of these uniquely identifiable datasets. The ESGF search provides a query approach to identify data subsets. This is underlying the data superset DOI concept. As a result, the ESGF as data disseminating infrastructure is prepared for the implementation of the RDA recommendations on evolving data.

Even the first step, the data superset DOI implementation, is a substantial improvement to the situation of CMIP5, where it was impossible to cite data. For CMIP6 data it will be possible to give credit to the data

creators for the first time in CMIP. Regarding the stability of CMIP6 data citations, the separation of the landing page URL from the data access makes the data superset DOIs stable against technology changes in the CMIP data infrastructure. Such changes require only a change in the database of the CMIP6 citation service. The long-term availability of CMIP data for access is granted by the IPCC-DDC. However, a long accessibility of CMIP data via the project pages can also be expected.[2]

What this first step of the CMIP6 citation implementation cannot establish, is a granted access to every dataset within the cited collection (single datasets might get unpublished and revised datasets might get ESGF published under new versions over time) and an automated identification of the used data subset.

3.1 Next Step: Citation of Used Data Subsets

The automated identification of the exact data subsets used across multiple CMIP6 citations was conceptually discussed with data publishing experts and CMIP6 data infrastructure developers resulting in a two PID approach integrating a data cart solution to mint data subset PIDs in addition to data superset DOIs. The use of a data cart PID selecting datasets over multiple CMIP6 data superset DOIs avoids the doubling of the number of data PIDs in a scholarly publication. In an intercomparison study comparing results from different models and experiments, the number of data superset citations can easily exceed 10 for model citations and 100 for experiment citations. Providing a single additional data subset PID on the data underlying a scholarly publication is more transparent for reviewers as well as readers of the article.

Current ESGF data cart implementation

The ESGF portal already offers such a data cart for the download of multiple datasets via automated script generation. The content of the data cart is stored in a table containing the IDs of the versioned datasets and the user IDs. A user has one cart on each ESGF index node (portal). Currently, data cart content is private, i.e. only visible for the specific user, and not shared among the index nodes.

The idea is to use this existing web form to add a possibility for users to share or publish their data cart contents and to request data subset PIDs for them. Such data subset PIDs are suitable to reference the data subset but also to exchange data cart contents among scientists. PIDs should be registered on those data carts, which are used in scholarly publications. Two steps are needed for a user to get a PID on the data cart:

- first publishing or sharing the cart with other users, and
- secondly requesting a PID for the public data cart.

It is highly unlikely that two users will use identical datasets, i.e. identical data cart content; however a scientist might want to share his/her data cart with others for collaboration.

ESGF data cart utilization for data subset PIDs

If the data cart sharing is heavily used by scientists, the storage of the data cart contents as individual dataset IDs gets expensive. Storing search queries in the data cart instead of the resulting individual dataset IDs provides an alternative. For each collection of datasets added to the data cart its underlying search query is stored.

The query approach has to be supported by the ESGF portals on the index:

- The carts need to be uniquely identifiable within the ESGF by cart IDs. The creation of several carts per user is required. Metadata exchange for public data carts among the index nodes should be aimed on the long-term.
- The storage of queries in the data cart has to be supported and enabled for a complete search result set as well as for the selection of individual datasets from the result set.
- Together with adding a query to the data cart the timestamp and a checksum, e.g. the checksum over the checksums of the files belonging to the datasets in the cart, should be added for verification.
- The portal or the index node has to provide a data cart page displaying its content without user login. For the query approach all queries in the data cart are executed with a logical or connection.

[2] CMIP3 data from 2005/2006 are still available from the project page at: https://esgf-node.llnl.gov/projects/cmip3/.

As data cart sharing turns private data carts into public ones, minimum provenance information need to be stored for verification. A user defined name or title of the data cart as reference is planned.

When a user wants to include a reference to a public data cart in a scholarly publication, a data subset PID on the data cart is registered upon user request. Data cart content can no longer be changed. The references of the data superset DOIs are added to the metadata. These are available from the citation service API (see Section 2.4 and **Figure 4**). The content of data carts with an assigned PID should be available on the long-term and thus replicated among the tier1 and tier2 nodes and a copy should be stored in a long-term archive such as the WDCC.

The combination of a data superset DOI on the full evolving data and a data subset PID on public data carts specifying the used subset of static ESGF dataset versions is compared to the RDA recommendations for evolving data citations (**Table 2**).

Comparison of the approach against the RDA recommendations

A. Preparing the data and query (R1–R3)

The individual datasets in the ESGF infrastructure are versioned and static. They are identifiable by their dataset IDs, which are short-hand notations of the dataset contents. Datasets are timestamped when published and unpublished. The dataset version string includes the date of publication. A faceted ESGF search functionality is available, on which query storage and query result display for the data carts are built, including a filter by version or ESGF publication date period.

B. Persistently identify specific datasets (R4–R10)

The additional storage of user queries for datasets in the data cart is planned. Every time a user adds datasets the underlying query is stored in this data cart.

ESGF provides a SHA256 checksums for the verification of individual files, which are part of versioned datasets. These can be used to verify the individual files in a downloaded result set. As files of unpublished datasets are missing in the result set of a data cart, a checksum for the stored queries is added, which is built on the checksums of the individual files or the file names (R6). The order of the downloaded files is not important. Therefore stable sorting is not relevant for this application (R5). Concerning data availability, the user can identify unpublished datasets on the landing page prior to download. For a result set verification, a verification service is planned, which calculates the individual SHA256 checksums of the downloaded files and the combined query checksum and compares individual and query checksums to the values stored in the metadata.

Stored queries of faceted ESGF search requests are normalized, having a fixed order of search attributes. Query execution results in a list of static datasets versioned by ESGF publication date, in which unpublished datasets are flagged as no longer available for download. The filter for ESGF publication date is part of the query. Thus additional query timestamping is not required (R7) but can provide useful provenance information on data cart content changes. As two identical data collections are highly unlikely, a check on query identity is not planned (R4).

An identifier is stored for a data cart (R8). The registration of a PID is planned upon user request. For these public data carts additional metadata on data citation is stored: author and title and references to superset DOIs for the data cart content (R10).

C. Upon Request of a PID (R11–R12)

The content of a public data cart is displayed on a landing page (R11). For data carts with registered PIDs citation information is added consisting of the data superset DOIs and the data cart PID.

Concerning R12, the machine actionability, the metadata is machine accessible using the ESGF search API. A fully automated download approach is not desirable because of the possibly high data volume.

D. Upon Modifications to the Data Infrastructure (R13–R14)

Technology changes (R13) require the transfer of data and all metadata, including dataset access. The transfer of the queries should not require additional effort apart from the metadata transfer. However, changes in the search facets caused by ESGF republication with different search facet names would require query rewrites and verification.

The combination of a data cart PID on the used data subset collection and the underlying data superset DOIs can be implemented as query approach along the RDA recommendations. If the query approach

Table 2: Comparison of the two PID approach for CMIP6 citation against the RDA-endorsed recommendations for the citation of evolving data.

RDA Recommendation	Extended CMIP6 Citation Concept (including Data Subset PIDs based on a data cart approach)
R1 – Data Versioning	Individual datasets in the data subset collection are versioned and static. The version is part of the dataset ID and part of the stored queries.
R2 – Timestamping	For the individual datasets in the data subset collection, the publication and unpublication is timestamped and the metadata is flagged. The time of publication is part of the version string. Versioned datasets are not changed. Changed datasets are published under new versions. Data cart content is not changed by data publication.
R3 – Query Store	The faceted ESGF search provides the query functionality. These queries are stored in the data carts. Individual queries include a filter by version i.e. ESGF publication period. Data cart content is reproduced by the execution of a combined query, connecting every stored query with a logical OR.
R4 – Query Uniqueness	The stored queries of the faceted ESGF search have a uniform order of search attributes. Because it is highly unlikely that two users will use identical data subsets in their data carts, query uniqueness is not checked.
R5 – Stable Sorting	Datasets as smallest subsets are static and consist of one or more individual files. The order of the downloaded files is not important. Record sorting is not relevant for this application.
R6 – Result Set Verification	The versioned datasets in the data cart are static and can be verified by their SHA256 checksums stored in the metadata. For the queries in the cart an additional checksum is available based on the result set in order to identify missing files in the downloaded data cart data.
R7 – Query Timestamping	Query results are static as their search results consist of static versioned individual datasets. Timestamping of the queries is not necessary but can provide useful provenance information on cart content history.
R8 – Query PID	Public data carts are assigned unique IDs, which are used for data cart content display. PIDs are registered upon user request on public data carts. Their contents are no longer changeable by the users.
R9 – Store Query	When the user adds data to a data cart, the query for the faceted ESGF search is stored in a normalized form together with the timestamp and a checksum. Additional metadata is stored for public data carts with registered PIDs, i.e. citation information and references to the data superset DOIs.
R10 – Citation Text	The citation recommendation is displayed on the landing page of the data subset PID including references to the data superset DOIs.
R11 – Landing Page	A landing page for the public data carts is provided. For public data carts with PIDs the citation recommendation is added.
R12 – Machine Actionability	A machine-readable version of the landing page based on the ESGF search API will be provided for public data carts including download information. Because of the possible high data volume, an automated download without checking the download volume beforehand, is not desirable.
R13 – Technology Migration	The query results are only dependent on the syntax of the faceted ESGF search. Technologically, a migration is a transfer of data cart metadata. Changes in the ESGF search facet names would require query rewrites and verification.
R14 – Migration Verification	not verified

will effectively reduce the required storage volume for data carts compared to the current storage of individual dataset IDs, is dependent on user behavior. If users add individual datasets to the data cart from a filtered result set, the query approach is not effective. In a first phase, the additional storage of queries together with the dataset IDs is planned. This provides a possibility to evaluate the query approach before replacing the current dataset ID approach for the data cart implementation of data subset PIDs.

3.2 Implementation Plan for CMIP6 and further Timeline

In December 2016 CMIP6 decided to postpone data subset citations to CMIP7 because of timeline and resource constraints.

The infrastructure component 'PID system' has developed a simple possibility to identify datasets by registration of a PID of type 'collection' on a list of objects baring PIDs. This implementation is limited to CMIP6 datasets and files with registered Handle IDs by the same system. The maximum number of datasets included is 100. The PID system and its services rely on copies of the ESGF's information on data identification and data access information stored in the Handle. By building their services on secondary metadata, an additional layer of complexity and possible failure is introduced. No policies on long-term data availability exist and no additional citation information is connected to these PIDs such as author or title etc., which would be needed to use them for data citation. In order to overcome the disconnection between these collection PIDs and data superset DOIs a currently developed CMIP6 registration service of scholarly publications using CMIP6 data will additionally collect data superset DOIs and data subset PIDs from the authors.

Within its limitations the PID system approach for a data subset PID can identify all files belonging to the data subset using the Handle metadata of the PID chain: data subset PID points to dataset PIDs point to the file PIDs, which have the data access information stored in their Handle metadata. These collection PIDs do not have queries stored explicitly, like the queries in the ESGF data carts. Thus a comparison to the RDA-endorsed recommendations for evolving data is not possible. The performance of this data subset PID implementation will be less than for queries on the ESGF index, esp. at times of massive PID registration or update of existing PID Handle metadata. However, the data subset PID approach is able to provide fine granular information on datasets underlying a scholarly publication within the CMIP6 project context. Thus statistics on dataset usage could be provided, e.g. on the variable usage of the data contributed by an institution, assuming that the authors register their scholarly publications together with the used collection PIDs at CMIP6.

Conclusion

The two PID concept consisting of a CMIP6 citation on the evolving data superset DOI and a deep citation of the used data portion with a data subset PID as data cart approach was presented. This approach could be implemented according to the RDA recommendations on evolving data. If a query approach is advantageous over the current storage of dataset IDs has to be proven in a first phase with additional storage of queries. For CMIP6 only the data superset DOIs will be implemented due to timeline and resource constraints. The implementation of the second part is postponed to CMIP7.

Compared to CMIP5, the provision of data citation information for CMIP6 data on the evolving data superset enables data users to give credit to the data creators. Tracing of data usage in scholarly publications is limited to the coarse data superset citation granularity. A collection PID provided by the new PID system adds a possibility for fine granular data identification under the precondition that the authors additionally register their scholarly publications at a separate service provided by CMIP6.

The tier1 and tier2 data nodes of ESGF are interested to publish own DataCite DOIs. The CMIP Citation Service is therefore likely to change from a central into a federated approach in the future. CMIP6 data superset DOIs can be transferred to a possible future ESGF member of DataCite.

The IPCC-DDC plans to add a data subset approach to their data citation service in future. The DDC currently provides a data cart for users similar to the ESGF. The implementation of a data cart approach as outlined for CMIP6 is much easier and straight forward for the DDC as data and metadata are stored at the DDC and the data is unchanging.

Acknowledgements

The authors would like to thank the colleagues at DKRZ, ESGF, WIP, and RDA, esp. the members of the RDA WGDC with Andreas Rauber and Martin Fenner for discussions around the citation of evolving data, as well as the Bundesministerium für Bildung und Forschung (BMBF; Federal Ministry of Education and Research) for funding (grant number: 01LP1605A).

Competing Interests

The authors have no competing interests to declare.

References

Altman, M, Borgman, C M and **Matone, M** 2015 An Introduction to the Joint Principles for Data Citation. *Bulletin of the Association for Information Science and Technology*, 41(3). Available at: http://www.asis.org/Bulletin/Feb-15/FebMar15_RDAP_Altman_EtAl.html [Last accessed 8 May 2017]. DOI: https://doi.org/10.1002/bult.2015.1720410313

ANDS 2017 Citing dynamic data. Available at: http://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation/citing-dynamic-data [Last accessed 8 May 2017].

ARGO 2017 Argo DOI, Digital Object Identifier. Available at: http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier [Last accessed 8 May 2017].

Austin, C, Bloom, T, Dallmeier-Tiessen, S, Khodiyar, V, Murphy, F, Nurnberger, A, Raymond, L, Stockhause, M, Tedds, J, Vardigan, M and **Whyte, A** 2016 Key components of data publishing: using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*. DOI: https://doi.org/10.1007/s00799-016-0178-2

Ball, A and **Duke, M** 2015 How to Cite Datasets and Link to Publications. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available at: http://www.dcc.ac.uk/resources/how-guides/cite-datasets [Last accessed 8 May 2017].

Brase, J, Irina, S I and **Lautenschlager, M** 2015 The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite. *D-Lib Magazine*, 21(1/2). DOI: https://doi.org/10.1045/january2015-brase

CMIP 2017 Coupled Model Intercomparison Project. Available at: http://cmip-pcmdi.llnl.gov/ [Last accessed 8 May 2017].

CMIP6 Citation Service 2017 Available at: http://cmip6cite.wdc-climate.de [Last accessed 8 May 2017].

DataCite Metadata Working Group 2016 DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. *Version 4.0. DataCite e.V.* DOI: http://doi.org/10.5438/0012

ES-DOC 2017 Earth System Documentation. Available at: http://es-doc.org/ [Last accessed 8 May 2017].

ESGF 2017 Earth System Grid Federation. Available at: http://esgf.llnl.gov/ [Last accessed 8 May 2017].

ESIP Stewardship Committee 2012 Parsons, M A, Barkstrom, B, Downs, R R, Duerr, R and Tilmes, C (eds.). *Data Citation Guidelines for Data Providers and Archives*. DOI: http://doi.org/10.7269/P34F1NNJ

Eyring, V, Bony, S, Meehl, G A, Senior, C, Stevens, B, Stouffer, R J and **Taylor, K E** 2016 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev.*, 9: 1937–1958. DOI: https://doi.org/10.5194/gmd-9-1937-2016

Fenner, M 2016 Dynamic Data Citation. In: Data Citation Working Group Mtg @ P8 Sep 16th 2016, Denver. https://rd-alliance.org/system/files/documents/160916_rda_p8_wgdc.pdf [Last accessed 8 May 2017].

FORCE11 Data Citation Synthesis Group 2014 Joint Declaration of Data Citation Principles. Martone, M (ed.). San Diego CA. Available at: http://www.force11.org/datacitation [Last accessed 8 May 2017].

FORCE11 FAIR Data Publishing Group 2017 FAIR Guiding Principles. https://www.force11.org/fairprinciples [Last accessed 8 May 2017].

International DOI Foundation (IDF) 2017 DOI Handbook. Available at: http://www.doi.org/doi_handbook/TOC.html [Last accessed 8 May 2017].

IPCC-DDC (Intergovernmental Panel on Climate Change Data Distribution Centre) 2017 Available at: http://ipcc-data.org/ [Last accessed 8 May 2017].

Klump, J, Huber, R and **Diepenbroek, M** 2016 DOI for geoscience data – how early practices shape present perceptions, *Earth Sci. Inform*. DOI: https://doi.org/10.1007/s12145-015-0231-5

Lawrence, B, Jones, C, Matthews, B, Pepler, S and **Callaghan, S** 2011 Citation and Peer Review of Data: Moving Towards Formal Data Publication, *International Journal of Digital Curation*, 6(2): 4–37. DOI: https://doi.org/10.2218/ijdc.v6i2.205

Mayernik, M, Callaghan, S, Leigh, R, Tedds, J and **Worley, S** 2015 Peer Review of Datasets: When, Why, and How. *Bull. Amer. Meteor. Soc.*, 96: 191–201. DOI: https://doi.org/10.1175/BAMS-D-13-00083.1

OpenAire 2017 OpenAire. https://www.openaire.eu/ [Last accessed 8 May 2017].

Orcid 2017 Connecting Research and Researchers. Available at: http://orcid.org [Last accessed 8 May 2017].

Parsons, M A and **Fox, P** 2013 Is data publication the right metaphor? *Data Sci. J.*, 12: WDS32–WDS46. DOI: https://doi.org/10.2481/dsj.WDS-042

Rauber, A, Asmi, A, van Uytvanck, D and **Pröll, S** 2015 Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). *Result of the RDA Data Citation WG*, 20 October 2015.

Available at: http://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf [Last accessed 8 May 2017].

Rauber, A, Asmi, A, van Uytvanck, D and **Pröll, S** 2016 Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bulletin of the IEEE Technical Committe on Digital Libraries*, 12(1). Available at: http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf [Last accessed 8 May 2017].

Scholix 2017 A Framework for Scholarly Link eXchange. Available at: http://www.scholix.org [Last accessed 8 May 2017].

Stockhause, M, Höck, H, Toussaint, F and **Lautenschlager, M** 2012 Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. *Geosci. Model Dev.*, 5: 1023–1032. DOI: https://doi.org/10.5194/gmd-5-1023-2012

Stockhause, M, Toussaint, F and **Lautenschlager, M** 2015 CMIP6 Data Citation and Long-Term Archival. *WIP White Paper*. Zenodo. DOI: https://doi.org/10.5281/ZENODO.35178

Taylor, K E, Juckes, M, Balaji, V, Cinquini, L, Denvil, S, Durack, P J, Elkington, M, Guilyardi, E, Kharin, S, Lautenschlager, M, Lawrence, B, Nadeau, D and **Stockhause, M** 2017 CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV's. version v6.2.3 (4 April 2017). https://docs.google.com/document/d/1h0r8RZr_f3-8egBMMh7aqLwy3snpD6_MrDz1q8n5XUk [Last accessed 8 May 2017].

WGCM-CMIP (WCRP Coupled Model Intercomparison Project) 2017 Available at: https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6 [Last accessed 8 May 2017].

WGCM Infrastructure Panel (Working Group on Climate Models Infrastructure Panel, WIP) 2017 Available at: http://www.earthsystemcog.org/projects/wip/ [Last accessed 8 May 2017].

How to cite this article: Stockhause, M and Lautenschlager, M 2017 CMIP6 Data Citation of Evolving Data. *Data Science Journal,* 16: 30, pp. 1–13, DOI: https://doi.org/10.5334/dsj-2017-030

Submitted: 17 October 2016 **Accepted:** 08 May 2017 **Published:** 15 June 2017

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS