

Early Citability of Data vs Peer-Review like Data Publishing Procedures

- Discussion Poster -

M. Stockhause^{1,2}, H. Höck¹, F. Toussaint¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ), ²Max Planck Institute for Meteorology (MPI-M)

EGU2014-3266



Motivation

The World Data Center for Climate (WDCC) has been assigning DOIs to long-term archived data of high quality for 10 years. Data becomes citable after the end of the scientific project. Modern data infrastructures like ESGF enable data sharing during the project phase, which has led to scientific publications based on datasets, which are still under revision. There are currently no common practices for such early citations of shared project data.

How to cite shared data before long-term archival?

Even for not yet long-term archived data, data users should be encouraged to cite the source data to give credit to the data creator(s). Agreements on requirements for early data citations as well as on the synchronization of data and article review procedures between data publishers and the mayor scientific publishers are essential for the acceptance of Early-DOI data citations by funding agencies and scientists.

Future Perspectives

The RDA/WDS Publishing Data IG brings together representatives from data and scientific publishers, DataCite, and the WDS to address practical aspects of the data publication concept. The problem of early data citations is a suitable topic for discussion within this IG. The long-term goal is the establishment of a research environment, which connects data with other data, with standard vocabularies, with scientific articles, and/or with a scientist's ID.

Initial Data Sharing Phase

Data is created, checked, and delivered by the creator. It is made accessible within a federated data infrastructure (e.g. Earth System Grid Federation) to registered users. Data analyses start and first scientific publications are submitted, normally by the data creating institute.

Force 11 Data Citation Principles

Minimal requirements for data citation principles were published in 02/2014 by Force11 in a joint declaration, wherein persistence is defined as the persistence of unique identifiers and metadata but not of the data.

Force 11 Joint declaration of data citation principles (02/2014) (<http://www.force11.org/datacitation>)

1. Importance
2. Scholarly Credit and Attribution
3. Evidence
4. Unique Identification
5. Access
6. Persistence: Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.
7. Specificity and Verifiability: Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. [...]
8. Interoperability and Flexibility

Early Data Publication / Citation:

The amount of data shared within the ESGF is typically large. Articles are based on multiple datasets. Data is to be cited as data collections in the reference lists.

For the developing data collections in this phase, persistent identifiers (e.g. DOIs) used for data citation point to a temporally changing aggregation of individual datasets.

Force 11's "Specificity and Verifiability" principle is not strictly fulfilled. Data Journals like ESSD with their additional requirement on the persistence of data, do not accept data papers based on such data.

Example for the early citation of data: NARCCAP / NCAR (<http://www.narccap.ucar.edu/about/citation.html>)

A DOI was assigned to the project data at the start of the data sharing phase, when the first datasets were accessible.

Citation Structure Recommendation:

Authors, publication year, year of last update. *Dataset Title*, Data Center. Download date. [Access path (URL or DOI)]
Additional full list of authors provided as well as acknowledgements.

Data Review Phase

Data is reviewed by the scientific community. Revised and new datasets are added as new versions, old datasets might be withdrawn. The overall quality of the data and the number of submitted papers increases and the rate of data revisions declines towards the end of the phase.

How to cite during the data review phase?

- Use unique persistent identifier (no authors, no title): suitable as reference in the text of scientific publications; not suitable for data creator's list of publications
- Use DataCite DOI incl. citation metadata: suitable for use in reference lists of scientific publications

How to distinguish Early-DOIs from LTA-DOIs?

- *By repository certifications?*
DataCite plans to include certificate information of its data publishers, e.g. WDS or DSA certificates; not sufficient for publishers of both kinds of DOIs
- *By different publisher names?*
Not yet stable data under review is published by under a second publisher name (like ESSDD for ESSD)
- *By a data quality flag?*
not supported by DataCite after asking its publishers
- *By different DOIs?*
suitable for different data collections of Early-DOI and LTA-DOI

Idea / Concept for Early-DOIs:

- Usage of different DOIs for Early- and LTA-DOIs
- Bidirectional connection of Early- and LTA-DOIs with specified relation types
- Guidance to use stable data by information and link to LTA-DOI on landing page of Early-DOI
- *Persistent access of Early-DOI data required?*

Requirements for the introduction of Early-DOIs:

- Agreement between *data and scientific publishers* on the *requirements and scope of application* for Early-DOIs, e.g. restricted to the review phase or restricted for use within a discussion paper.
- Agreement between *data and scientific publishers* on the synchronization of data and paper review procedures in a similar way as for embargoed data.
- Acceptance of the synchronization of data and paper review processes by *authors* as the final paper publication might be delayed by the LTA data publication.

Stable Data Phase

Data is long-term archived, normally at the end of the research project. Data reuse and the submission of scientific publications continues. The data publication of long-term archived (LTA) data is well established, LTA data references are widely accepted by scientific publishers.

Data Journal Requirements for Repositories

To grant permanent access to the published data, data journals have higher minimal requirements for data repositories than Force 11, especially on access and persistence of the data.

Earth System Science Data (www.earth-system-science-data.net)

Certified repositories (case to case decision)

Persistent Identifier, e.g. DOI
Open Access: free of charge without any barriers

Liberal Copyright: equivalent to the Creative Commons Attribution License

Long-term Preservation: guarantee a long-term availability of the data sets and a **permanent access**.

Geoscience Data Journal (www.geosciencejournal.com)

Registered repositories at GDJ (case to case decision)

Persistent Identifier: DOI
Open Access: to the broad/public user community; easily readable formats; supporting metadata

Liberal Copyright: -- (GDJ promotes the sharing and re-use of research data)

Long-term Preservation: as long as the data is **scientifically beneficial**.

LTA Data Publication / Citation at WDCC

(<http://www.dkrz.de/daten-en/Datapublication>)

WDCC is certified by WDS. The DOI data publication process is restricted to long-term archived data. It includes a thorough quality self-assessment (QA) by the data manager (technical QA) and the data creator (scientific QA).

For large data collections like for CMIP5 this DOI data publication process is finished after the project report is written, in case of CMIP5 the IPCC AR5 part 1. Data creators do not get their earned credit.

Example for the late citation of data: CMIP5, CORDEX / WDCC (e.g. <http://verc.enes.org/data/projects/cordex/data-citation/data-citation>)

DOIs are assigned to data collections of one or a couple of connected model simulations after quality assessment and long-term archiving of the data (after the end of the project).

Citation Structure Recommendation:

Authors, publication year. *Dataset Title*, Data Center. DOI.

Initial data accessibility

Initial 1st data version available; Data analyses

Back-up of shared data

Strictly versioned data added and withdrawn; Data analyses by additional users

Long-term archival (LTA) of data

Data unchanged; Extensive data reuse

Data Workflow

First papers written based on early data

Draft paper and early data publication

Paper and data review

Final paper and LTA data publication

Publications based on LTA data

Publication Workflow



stockhause@dkrz.de
WDC Climate: www.wdc-climate.de
DKRZ: www.dkrz.de

