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.

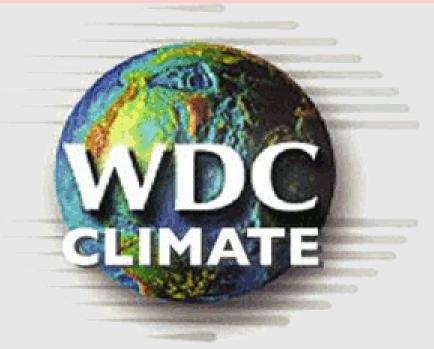
Initial Data Sharing Phase

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.	Early Data Publication / Citat The amount of data shared wit large. Articles are based on mu to be cited as data collections i For the developing data collect persistent identifiers (e.g. DOIs point to a temporally changing datasets.
	 Force 11 Joint declaration of data citation principles (02/2014) (http://www.force11.org/datacitation) 1. Importance 2. Scholarly Credit and Attribution 3. Evidence 	Force 11's "Specificity and Veri strictly fulfilled. Data Journals I additional requirement on the p accept data papers based on s
	 4. Unique Identification 5. Access 6. Persistence: 	Example for the early citation of dat (http://www.narccap.ucar.edu/about/cit
	Unique identifiers, and metadata describing the data, and its disposition, should persist even beyond the lifespan of the data	A DOI was assigned to the project data sharing phase, when the first datasets
	they describe. 7. Specificity and Verifiability: Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. []	<i>Citation Structure Recommendation</i> Authors, publication year, year of last u Data Center. Download date. [Access
	8. Interoperability and Flexibility	Additional full list of authors provided a

Initial data accessibility Initial 1st data version available; Data analyses

First papers written based on early data



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)

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.

tion:

ithin the ESGF is typically ultiple datasets. Data is in the reference lists.

tions in this phase, s) used for data citation aggregation of individual

rifiability" principle is not like ESSD with their persistence of data, do not such data.

ata: NARCCAP / NCAR itation.html)

ta at the start of the data s were accessible.

update. Dataset Title, path (URL or DOI)]

list of authors provided as well as acknowledgements.

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

Back-up of shared data

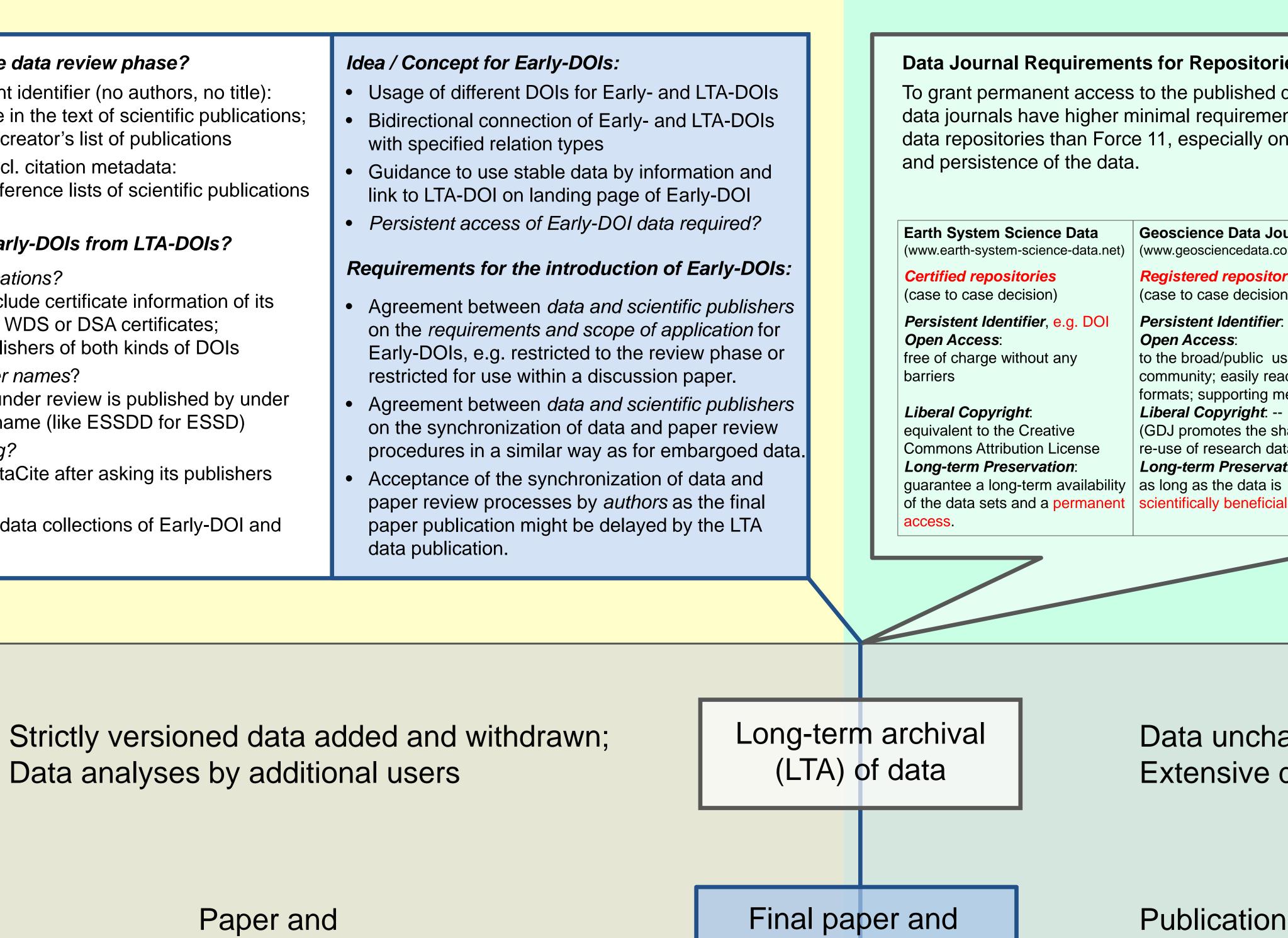
Data analyses by additional users

Draft paper and early data publication

Paper and data review

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

Data Review Phase



LTA data publication

EGU2014-3266



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).

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 acces

> Geoscience Data Journal (www.geosciencedata.com) **Registered repositories** at GE (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: 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/datacitation)

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

Citation Structure Recommendation: Authors, publication year. Dataset Title,

Data Center. DOI.

Data unchanged; Extensive data reuse

Publications based on LTA data



Publication Workflow



KLIMARECHENZENTRUM