

# CMIP6 Citation Service - Review and Perspectives

Martina Stockhause<sup>1,1</sup>

<sup>1</sup>German Climate Computing Centre (DKRZ)

November 30, 2022

## Abstract

Data publication with DOI assignment has become common practice. The Citation Service for Coupled Model Intercomparison Project Phase 6 (CMIP6) was requested by the scientific panel WGCM, which is part of the World Climate Research Programme (WCRP) to enable references of the CMIP6 data in the upcoming Sixth IPCC Assessment Report (AR6). Expectations for data citations include aspects of data usage metrics to receive credit and reproducibility of published research findings. These two are difficult to combine, because a meaningful data usage metrics requires information on large data collections such as experiment data while reproducibility relies on individual datasets. In addition, data usage metrics rely on data users citing the data in the reference list of a scholarly publication and the publisher to provide data references in the articles' metadata. Neither is a given. Organizations like the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) play an important role for the required ongoing community change. The Reliquaries approach together with PID Graph implementations being developed in the Data Citation Community of Practice subgroup shows promise for combining reproducibility and credit expectations. The contribution will present survey results on CMIP6 participants' expectations, discuss gaps in the current citation system and investigate, how new ideas can help to close these gaps. Reference: Stockhause, M. and Lautenschlager, M., 2017. CMIP6 Data Citation of Evolving Data. Data Science Journal, 16, p.30. DOI: <http://doi.org/10.5334/dsj-2017-030>

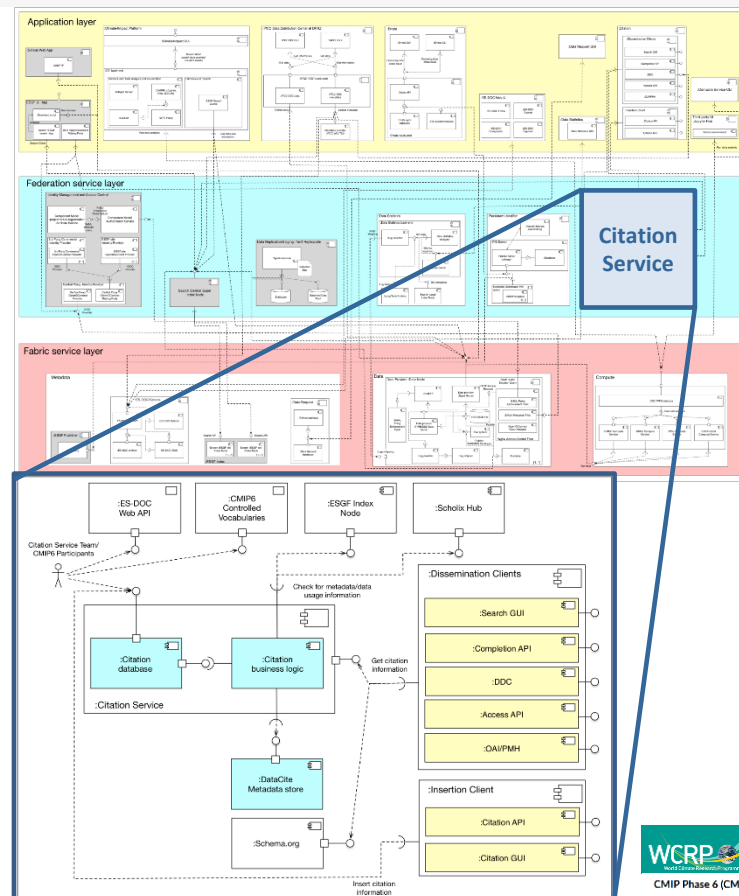
# CMIP6 Citation Service: Review and Perspectives

Citation service as part of the decentral federated CMIP6 infrastructure

Coupled Model Intercomparison Project Phase 6 (CMIP6) is a project of the World Climate Research Programme (WCRP), which is jointly coordinated by CMIP and WGCM Infrastructure Panels.

Challenge of supporting data providers, infrastructure partners and data users with dedicated services for human as well as machine-access.

The citation service was very successful:  
All CMIP6 datasets were citable on the AR6 WGI cut-off date  
2021-01-31.



<http://cmip6cite.wdc-climate.de>

[http://bit.ly/CMIP6\\_Citation\\_Search](http://bit.ly/CMIP6_Citation_Search)

Stockhouse and Lautenschlager., 2017. <http://doi.org/10.5334/dsj-2017-030>

# CMIP6 Citation Service: Review and Perspectives

The traditional citation use cases are complemented by data services reusing data.

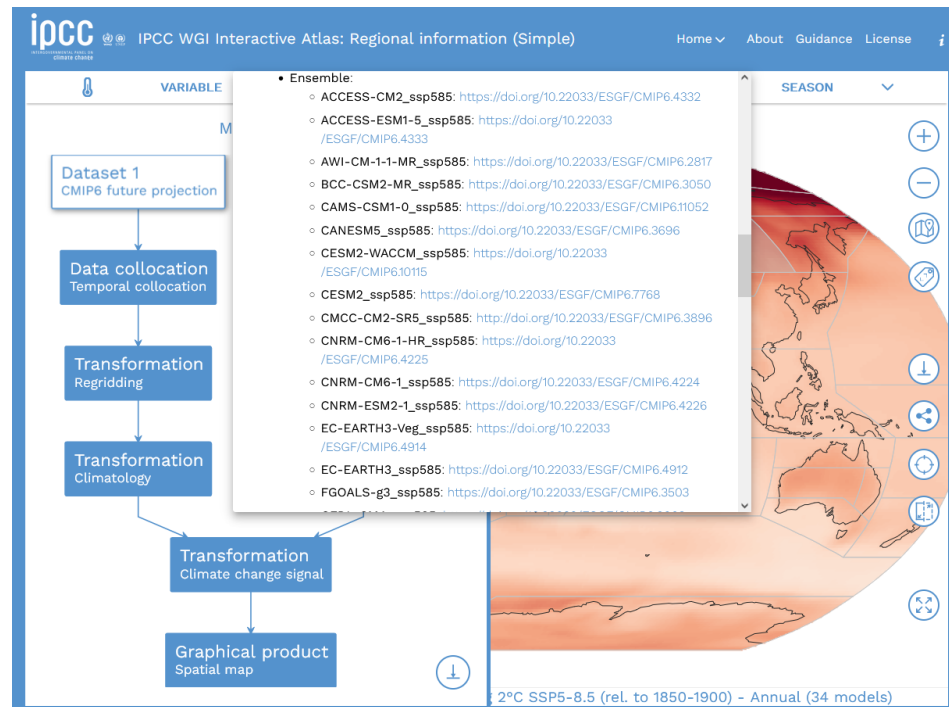
## Traditional Usage:

- Data Citation in (static) scholarly publications

## New Usages:

- “Dark archives” of CMIP6 data copies
- Climate Services providing domain specific (dynamic) data products
- Platforms providing access to multiple data providers, e.g. Copernicus Climate Data Store
- Cloud Services, e.g. Pangeo
- AI applications

Few services comply to the CC BY license requirement and cite the CMIP6 data, e.g. IPCC WGI Interactive Atlas



<https://interactive-atlas.ipcc.ch/>

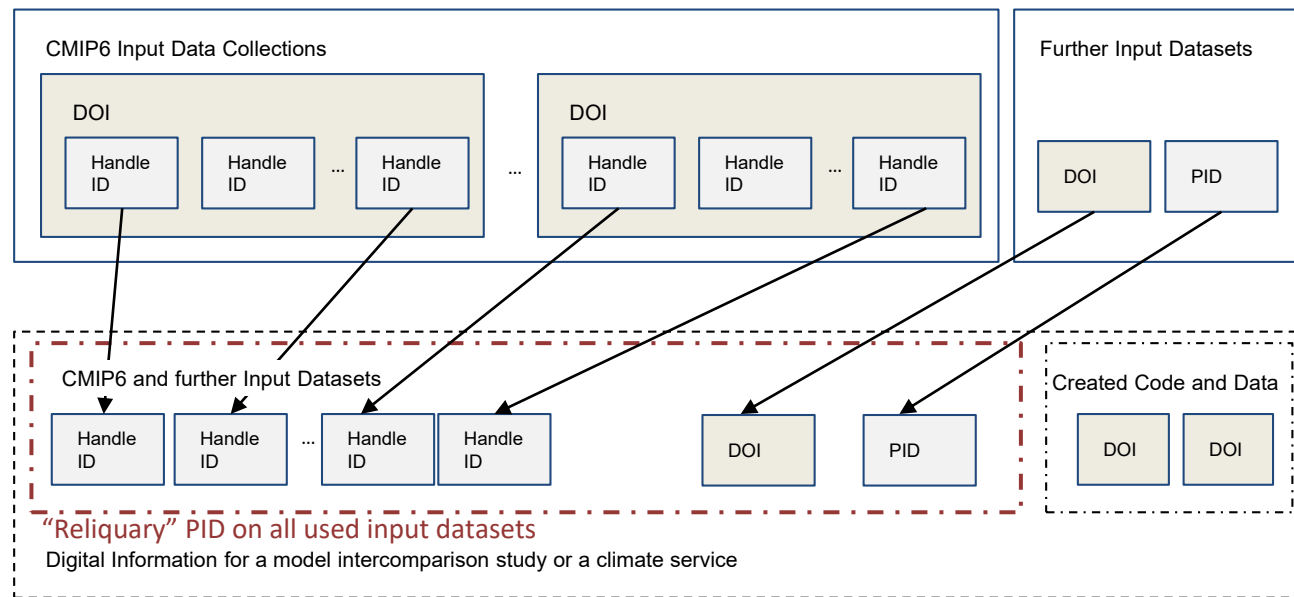
# CMIP6 Citation Service: Review and Perspectives

Complex citation cases and the “Reliquary” PID concept support data citation and data reproducibility at the same time

A “Reliquary” PID on all input datasets used in a model intercomparison study or a climate service is suggested.

Technical requirements are:

- Cross-repository PID assignment
- Credit needs to be directed to the “Reliquary” content providers and not to the “Reliquary” creator.



CMIP6 Citations are one piece in the interoperable FAIR Digital Object framework of the future.

# CMIP6 Citation Service - Review and Perspectives -

AGU Fall Meeting 2021, 2021-12-17

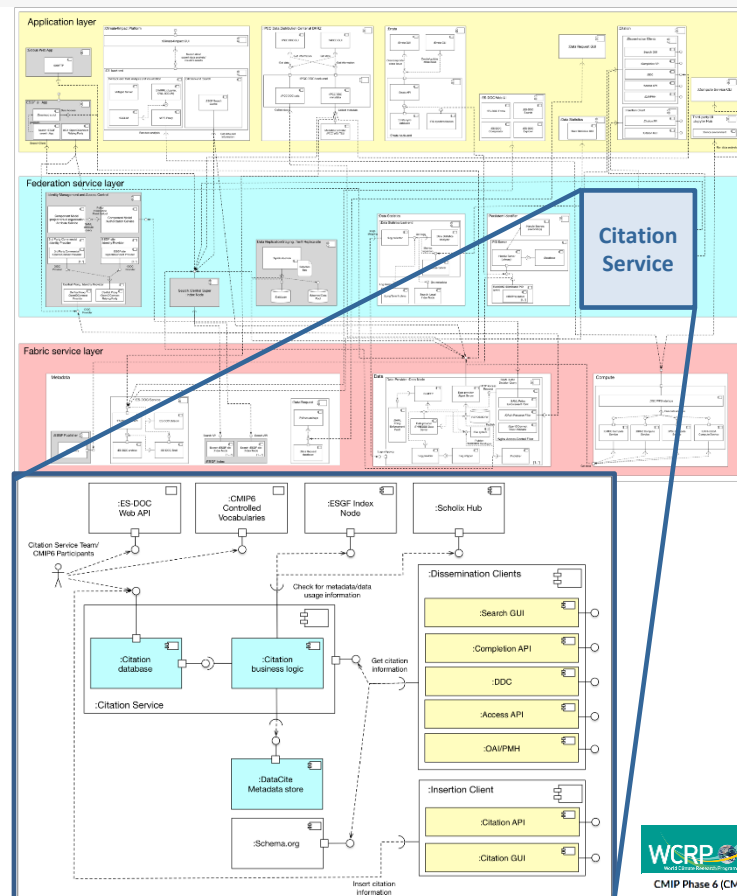
Martina Stockhause (DKRZ/IPCC DDC)

# 0. Introduction to the CMIP6 Citation Service

## Citation service as part of the decentral federated CMIP6 infrastructure

- Coupled Model Intercomparison Project Phase 6 (CMIP6) is a project of the World Climate Research Programme (WCRP), which is jointly coordinated by CMIP and WGCM Infrastructure Panels.
- Citations are strongly related to both - data and documentation.
- Citation Service has to support data providers, infrastructure partners and data users.
- Citation Service has to serve human- and machine-access.

Challenge of integration into different systems (incl. third party systems) and therefore maintenance of multiple interfaces



<http://cmip6cite.wdc-climate.de>

[http://bit.ly/CMIP6\\_Citation\\_Search](http://bit.ly/CMIP6_Citation_Search)

Stockhouse and Lautenschlager., 2017. <http://doi.org/10.5334/dsj-2017-030>

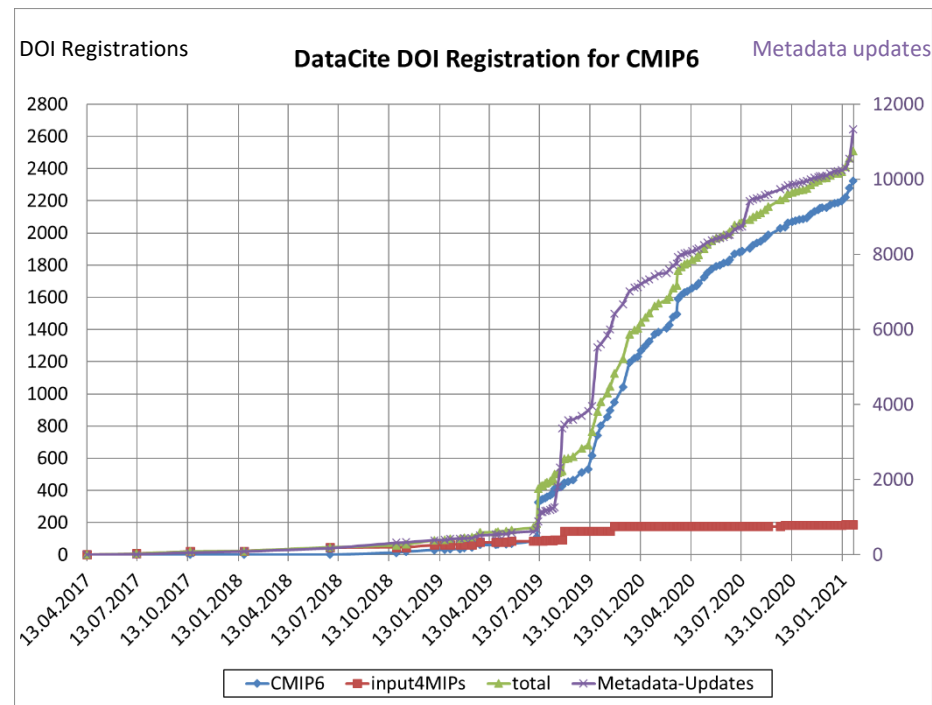
# 0. Introduction to the CMIP6 Citation Service

The citation service was very successful: All CMIP6 datasets were citable on the AR6 WGI cut-off date 2021-01-31.

- Data providers maintain citation content (DataCite content) through a GUI and an API
- Automated services register DOIs for new ESGF experiment data, curate metadata, and monitor the service
- Data users are provided different APIs and a graphical search interface to find data citations; data citations can be harvested from an OAI server or crawled from the landing pages (schema.org standard)

## DOI Status on 2021-01-31:

Number of DOIs:	2 509
• CMIP6:	2 322
• input4MIPs:	187
Number of MD updates:	13 876
Number of Scholix references:	63



[http://bit.ly/CMIP6 DOI Statistics](http://bit.ly/CMIP6_DOI_Statistics)

# 1. Importance of the CMIP6 Citation Service

Data citation become part of scholarly workflows and an obligation due to data license assignments.

For scientists/data creators (survey):

- Receive credit for data, DOI is the desired PID type
- Data Impact matrix with special importance of citations in scientific literature and the IPCC AR6

For funders:

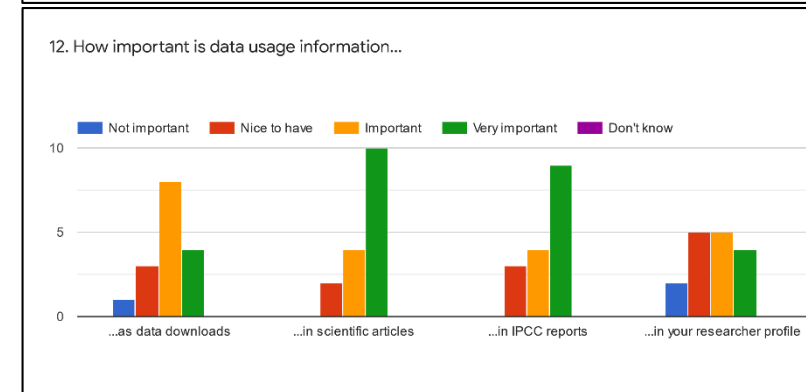
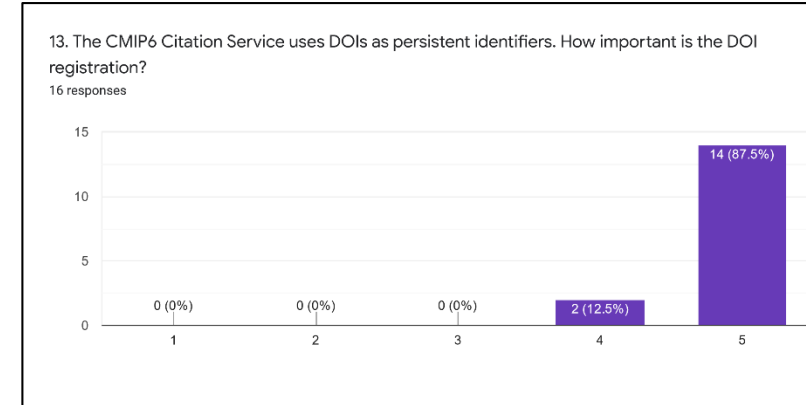
- Receive credit for provided funding grant

For data users:

- Give credit for used input data
- Comply to author guidelines

For data using organizations/companies/service providers:

- Enhance transparency of results, e.g. IPCC's FAIR Guidelines
- Comply to the CC-BY license requirement to give credit by data citation



<http://doi.org/10.5281/zenodo.5534136>

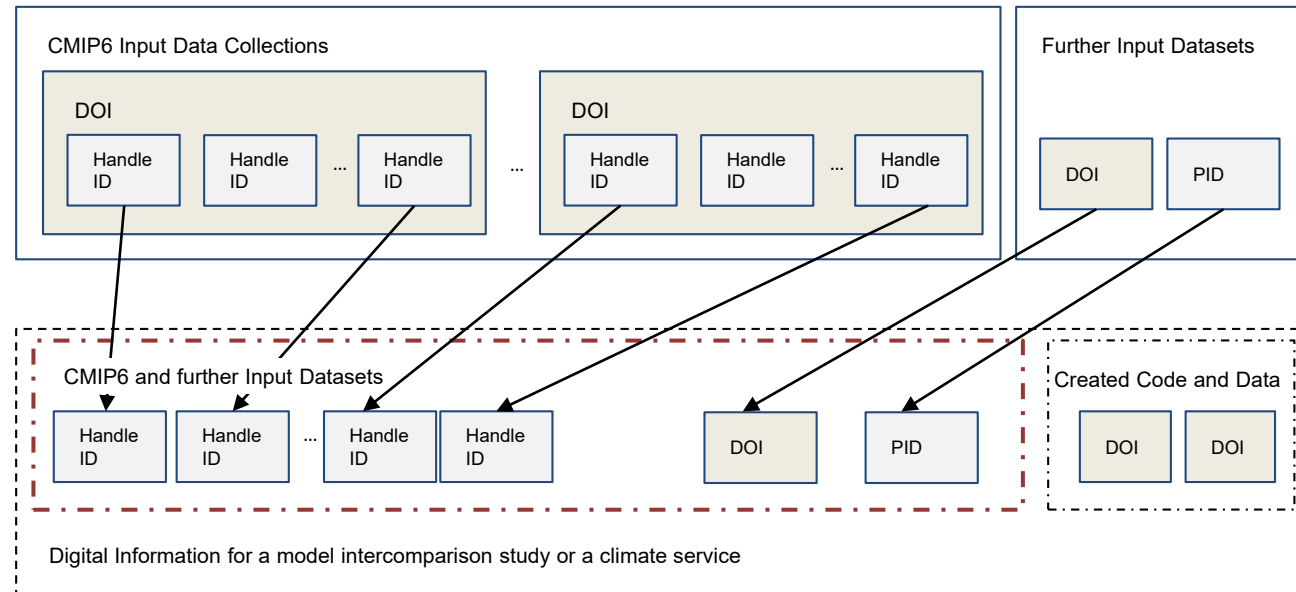


# 1. Importance of the CMIP6 Citation Service

Credit is not sufficient for reproducibility for high-volume data (cited on data collections)

- Data citation is provided on experiments (many datasets created together)
- Data access and usage granularity is a dataset
- Model intercomparison studies or climate services use few datasets across several experiments

Reproducibility requires the identification of individual datasets used in a study in addition to data citation.



## 2. Usage of the CMIP6 Citation Service

The traditional citation use cases are complemented by data services reusing data.

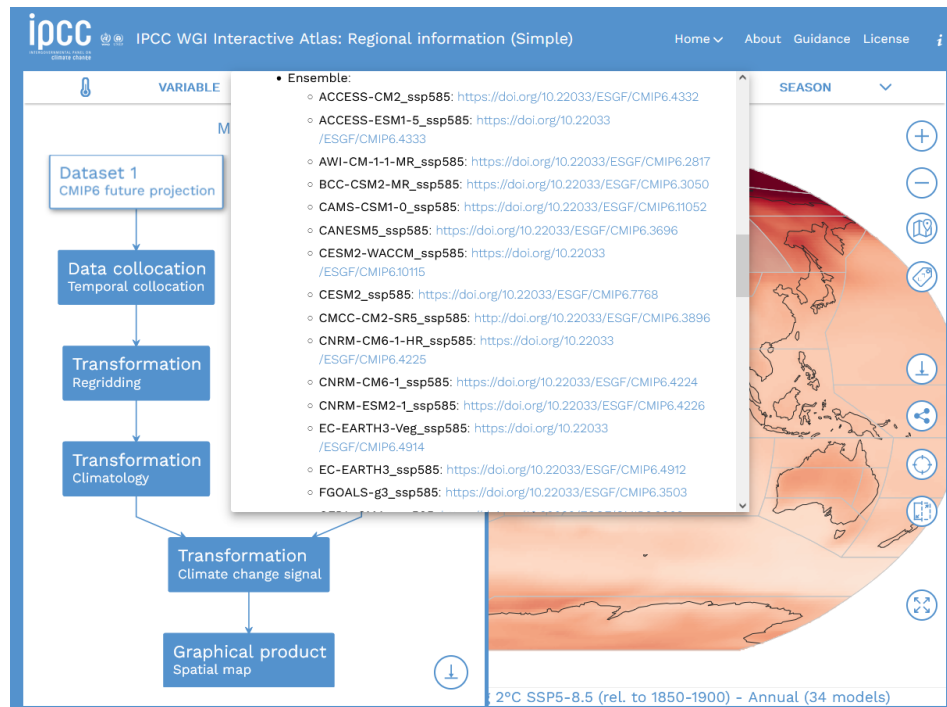
### Traditional Usage:

- Data Citation in (static) scholarly publications

### New Usages:

- “Dark archives” of CMIP6 data copies
- Climate Services providing domain specific (dynamic) data products
- Platforms providing access to multiple data providers, e.g. Copernicus Climate Data Store
- Cloud Services working with data converted into zarr format, e.g. Pangeo
- AI applications

Few services comply to the CC BY license requirement and cite the CMIP6 data, e.g. IPCC WGI Interactive Atlas

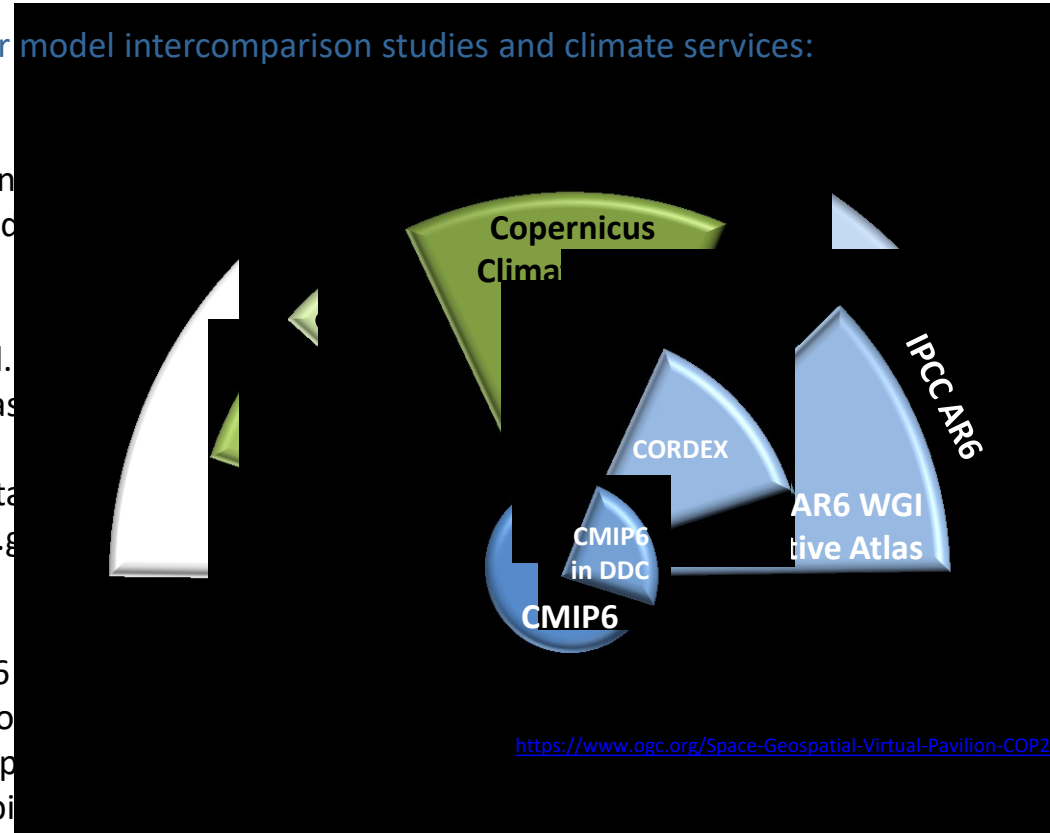


<https://interactive-atlas.ipcc.ch/>

## 2. Usage of the CMIP6 Citation Service

IPCC FAIR Guidelines as best practice example for model intercomparison studies and climate services:

- **Credit/License**
  - Credit creators of input datasets like CMIP6 by citing the data in publication and climate services based on CMIP6 data
  - CC BY-SA license requires credit
- **Transparency/Trust**
  - Documentation of input datasets (incl. data citations), scripts and final data as provenance information
  - Publish DOI references to related digital objects to make them discoverable, e.g. by PID graphs
- **Preservation**
  - Political decisions are based on CMIP6 Curation and long-term preservation of datasets and main infrastructure components is required for their long-term traceability



Preservation is the most challenging aim, which requires a long-term funding strategy of the core components.

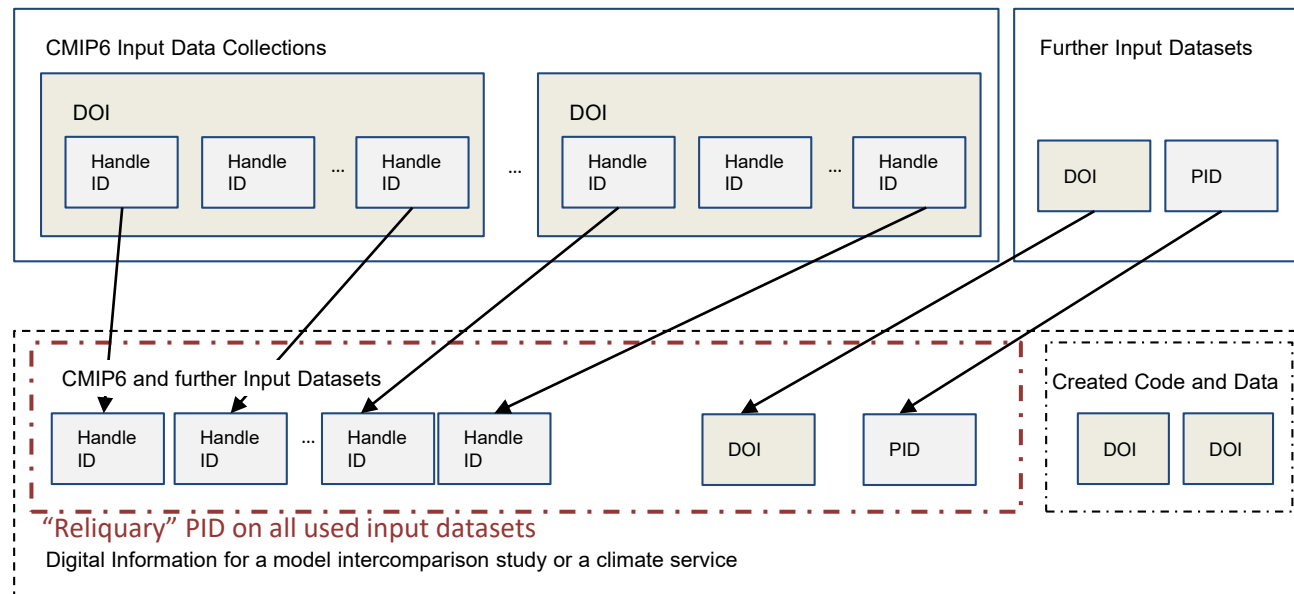
### 3. Future Perspectives

#### Complex citation cases and the “Reliquary” PID concept

A “Reliquary” PID on all input datasets used in a model intercomparison study or a climate service enables reproducibility and data citation at the same time.

#### Technical requirements are:


- Cross-repository PID assignment, which is easy to implement in analysis tools and easy to use by scientists
- Credit needs to be directed to the “Reliquary” content providers and not to the “Reliquary” creator.



CMIP6 Citations are one piece in the interoperable  
FAIR Digital Object framework of the future.

<https://cmip6cite.wdc-climate.de>

stockhause@dkrz.de

 [orcid.org/0000-0001-6636-4972](https://orcid.org/0000-0001-6636-4972)