Data Publication Process for CMIP5 Data and the Role of PIDs within Federated Earth System Science Projects

M. Stockhause[1,2], H. Höck[1], F. Toussaint[1], T. Weigel[1,3], and M. Lautenschlager[1]

[1]German Climate Computing Center (DKRZ), [2]Max Planck Institute for Meteorology (MPI-M), [3]Universität Hamburg

AGU 2012: IN23C-1525

Motivation

CMIP5[1] data are published on worldwide distributed data nodes of a grid infrastructure, the ESGF[2]. Additional replica are held by larger data centres. Therefore data consistency is of special importance. It relies on unique identification of data.

The experience of CMIP5 and the analysis of identifiers' usage within CMIP5 has motivated the development of a PID concept for federated infrastructures in Earth System Sciences.

Identification of Data and Metadata in CMIP5

For CMIP5 ca. 3 PB of data are expected to be published at ESGF data nodes worldwide. An ESGF dataset is a collection of files. For the long-term archived data a DOI is assigned on a collection of ESGF datasets. CMIP5 deals with 3 different data aggregation levels.

Apart from data, independent metadata are created in the CIM questionnaire and during the QC process. All information is harvested by the gateways and available for data discovery.

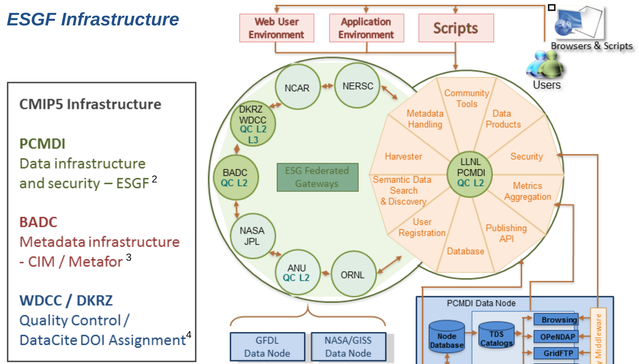
Future Perspective with PIDs

Distributed and federated data infrastructures are important in *'big data'* scientific communities such as Earth System Sciences. For a unique identification of files and pieces of information, reliable and unique identifiers are indispensible.

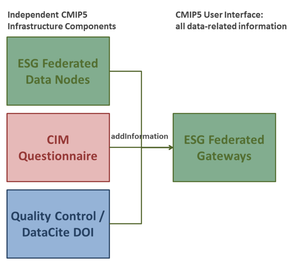
PIDs with handles, which store relations between pieces of data and information along with the location (URL), can provide such identification, relation, and provenance information.

CMIP5 Infrastructure and Technical Workflow

ESGF Infrastructure



CMIP5 Technical Workflow



CMIP5 Infrastructure Aspects

The ESGF data infrastructure[2] was supplemented by a CIM questionnaire[3] and a quality control / DataCite DOI service[4]. These independent infrastructure components provide additional information on model, experiment design, simulation, quality, and citation.

In the ESGF[2] gateways, these different pieces of data-related information are collected and merged. Then they are available for data discovery. The DRS_ids (Data Reference Syntax Identifier) are used for the identification of relations among them. Each infrastructure component has defined internal identifiers for its datasets (i.e. collections of data).

More Information:

[1] *http://cmip-pcmdi.llnl.gov/cmip5/*
[2] *http://esgf.org*
[3] *http://metaforclimate.eu*
[4] *http://cmip5qc.wdc-climate.de*

Identifiers Used within CMIP5

Comparison of the different Identifiers used within CMIP5

Name of Identifier	Describes	Creation Process	Usage within CMIP5 By	Useable For	Uniqueness	Consistency
DRS_id (ESGF)	• files • ESGF datasets • DataCite datasets • any other data collection	CMOR2 writes the directory structure into the netCDF file header as DRS-CMOR2 components; ESGF publication creates ESGF DRS_ids (data version added) .	ESGF gateways harvest all information; ESGF data nodes down to ESG datasets; CIM for simulations; QC down to files.	• Matching between data node, CIM, and QC information consistency checks with mapping effort	DRS-CMOR2 is not unique (no version); DRS-ESGF is unique but used partly with deviations in component names and structure (e.g. simulation definitions differ between data/QC and CIM).	Deviations occur for CIM, and in the ESGF data nodes for files. DRS exists in 2 versions: CMOR2 and ESGF. No controlled vocabulary is used for the creation of the DRS components.
tracking_id	• files	CMOR2 writes tracking_ids in the netCDF file header. no 1:1 relation to ESGF DRS_ids	Published by ESGF data nodes; Used for QC consistency checks	• Data creation process consistency checks	Not always, because its creation is not enforced	Inconsistencies occur when data changes are performed outside CMOR2.
MD5 checksum	• files	The checksum is created by the ESGF data node manager prior to publication.	Used by ESGF gateways during data replication. Used for QC consistency checks	• Control of data transfer • consistency checks	Yes, but not enforced to be created prior to ESGF publication.	Inconsistencies occur when changed data is published under a new version without checksums update.
DataCite DOI	• DataCite datasets	DOIs are assigned by the Publication Agency at the end of the QC process when the data is archived.	Published and used by the Publication Agency, i.e. in CMIP5 WDCC.	• Useable in scientific publications as data reference	Yes with persistent data access.	Yes
CIM documentID	• CIM documents	CIM creates them for new CIM documents.	Internal usage within CIM	• Only within CIM	Within CIM in combination with version	Consistency within CIM most likely

Identifiers used within CMIP5

In CMIP5 three data identifiers are used within the infrastructure:

- DRS_id
- tracking_id
- MD5 checksum

The DRS_id is the only identifier, which can be used for data collections. Therefore it is the one used for matching CIM and QC information to the data.

The DataCite DOI is assigned to the long-term archived CMIP5 data / DDC[5] data. It is used to cite the data in scientific publications.

MD5 checksums were introduced to check replicated data for consistency.

More Information:

[5] *http://www.ipcc-data.org/*

Experience with CMIP5 Identifiers

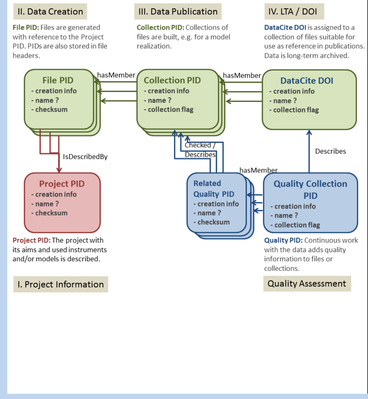
- Consistent usage of identifiers is not enforced but depend on the usage of specific tools or the data node managers. E.g. an inconsistent versioning leads to inconsistencies between original data and replica.
- DRS_id components have no controlled vocabulary but are constructed during ESGF data publication. DRS_ids are only strictly used down to the ESGF dataset level. Thus, mapping efforts are required for matching between data, CIM, and QC.

The DRS_id for data and data collection identification together with the MD5 checksums for consistency checks would be sufficient in a central data infrastructure. For a federated data infrastructure a consistent usage of these identifiers cannot be fully achieved.

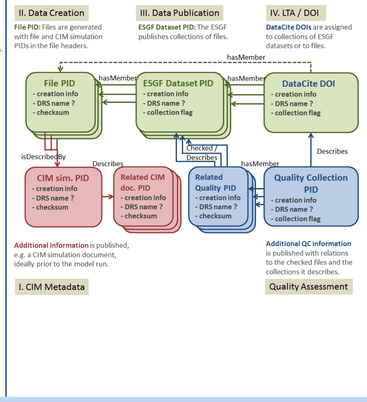
Stockhause et al. (2012), Geosci. Model Dev., DOI:10.5194/gmd-5-1023-2012 .

PIDs in Federated Earth System Science Projects

General PID Concept



ESS Example



PID Concept

Data and information are assigned PIDs for individual file as well as for collections. Relations among these are recorded in Handle[6] key-metadata. A name key can provide the possibility to assign a relation, manually. A data collection needs to be closed because of the relations to other PIDs. Metadata PIDs can be added later, e.g. a quality PID.

A typical workflow within a project is:

1. Project definition (aims, work packages,…)
2. Data creation
3. Data collection publication
4. Long-term archiving / DataCite[7] DOI assignment

Quality information is added, continuously.

ESS Example with PID application

In Earth System Sciences (ESS) data is created once and used multiple times. Often by users, which were not involved in data production. Therefore detailed metadata about the scientific project as well as about the data creation process need to be collected along with the data.

As metadata is often provided after data creation. The introduction of a data name key could provide for a subsequent manual definitions of data-metadata relations.

The success of the PID concept requires central tools such as CMOR, the ESGF data node and CIM to assign PIDs. In addition, the support of DataCite[7] is essential.

More Information:

http://rd-alliance.org
[6] *http://handle.net*
[7] *http://datacite.org*

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

DEUTSCHES KLIMARECHENZENTRUM