

Data Publication and Quality Control Procedure for CMIP5 / IPCC-AR5 Data

Martina Stockhause, Michael Lautenschlager, Heinke Hoeck, and Frank Toussaint

EGU2011-2859

CMIP5 Organization & Infrastructure Components

For distribution of data connected to the next IPCC report, the Earth System Grid Federation (ESGF) was founded. Its members have different responsibilities within the data infrastructure:

- **PCMDI / LLNL:** data and security infrastructure (ESG)
- **BADC (British Atmospheric Data Centre):** metadata infrastructure (METAFOR / CIM)
- **WDCC (World Data Center Climate) / DKRZ:** quality control, data publication (DataCite DOI)

CMIP5 Quality Control (QC)

For CMIP5 ca. 3 PB of officially requested data are expected to be archived. About 1 PB of that data will likely be of especially high interest and will be replicated by the three ESGF partners.

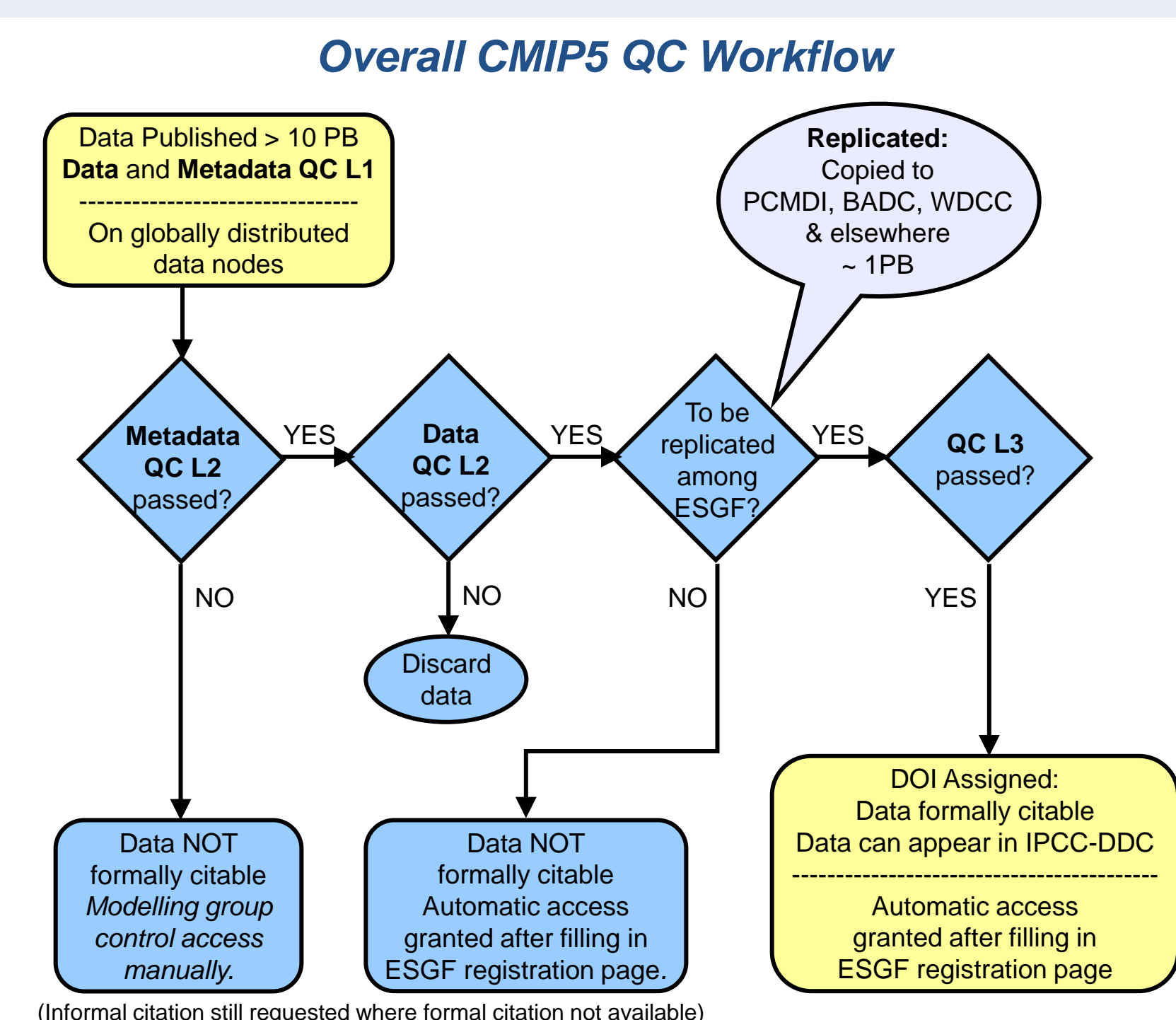
Because of the high data volume the QC checks up to level 2 are performed distributed among the ESGF. The final QC level 3 checks for the DOI assignment are carried out by WDCC. Afterwards CMIP5 data is formally citable and remains persistent.

Future Perspective

The current DOI publication procedure is comparable to the publication of grey literature in scientific print media. For the integration of a peer review process quality procedures accepted and agreed on by the earth system modelling community are necessary.

The distributed quality control approach could be reduced in complexity by the integration of the QC Repository into the CIM Metadata Repository.

CMIP5 Quality Control Workflow



Quality Levels for CMIP5

Three Quality Control (QC) Levels are defined for CMIP5 data:

- **QC Level 1:** **Metadata:** Technical checks on METAFOR questionnaire input data
Data: CMOR2 and ESG publisher conformance checks
- **QC Level 2:** **Metadata:** METAFOR questionnaire metadata checked by scientist
Data: Technical checks e.g. on the reliability of variable ranges and the consistency checks between data and data requirements

QC Level 3 / DOI:

Data approved by author and published as DOI

QC checks for data and metadata are performed, separately, for levels 1 and 2. During the cross-checks of QC L3 checks their results are reviewed.

Data assigned a DOI is formally citable and is granted persistent access.

Quality Control Workflow CMIP5

The different QC levels are connected with different access rights for registered users:

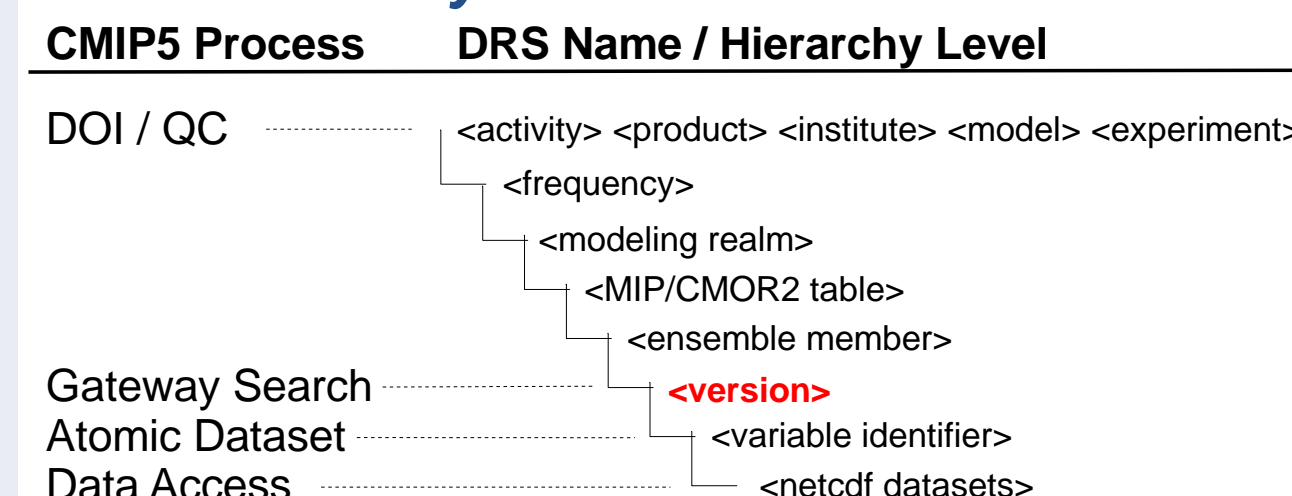
- **Restricted Access (QC Level 1 Data):** After ESG publication the access is restricted under control of the modelling centre.
- **Scientific Access (QC Level 2):** The scientific community is granted access of data of QC L2.
- **QC Level 3 / DOI:** With the DOI assignment the data archive is opened for access by non-scientific users.

Granularity of Quality Control

QC is accomplished on DRS Atomic Dataset level. The QC results are aggregated on DRS experiment level.

In the gateways data discovery is supported down to the level of Ensemble versions (ESG dataset).

Granularity of QC in CMIP5 context

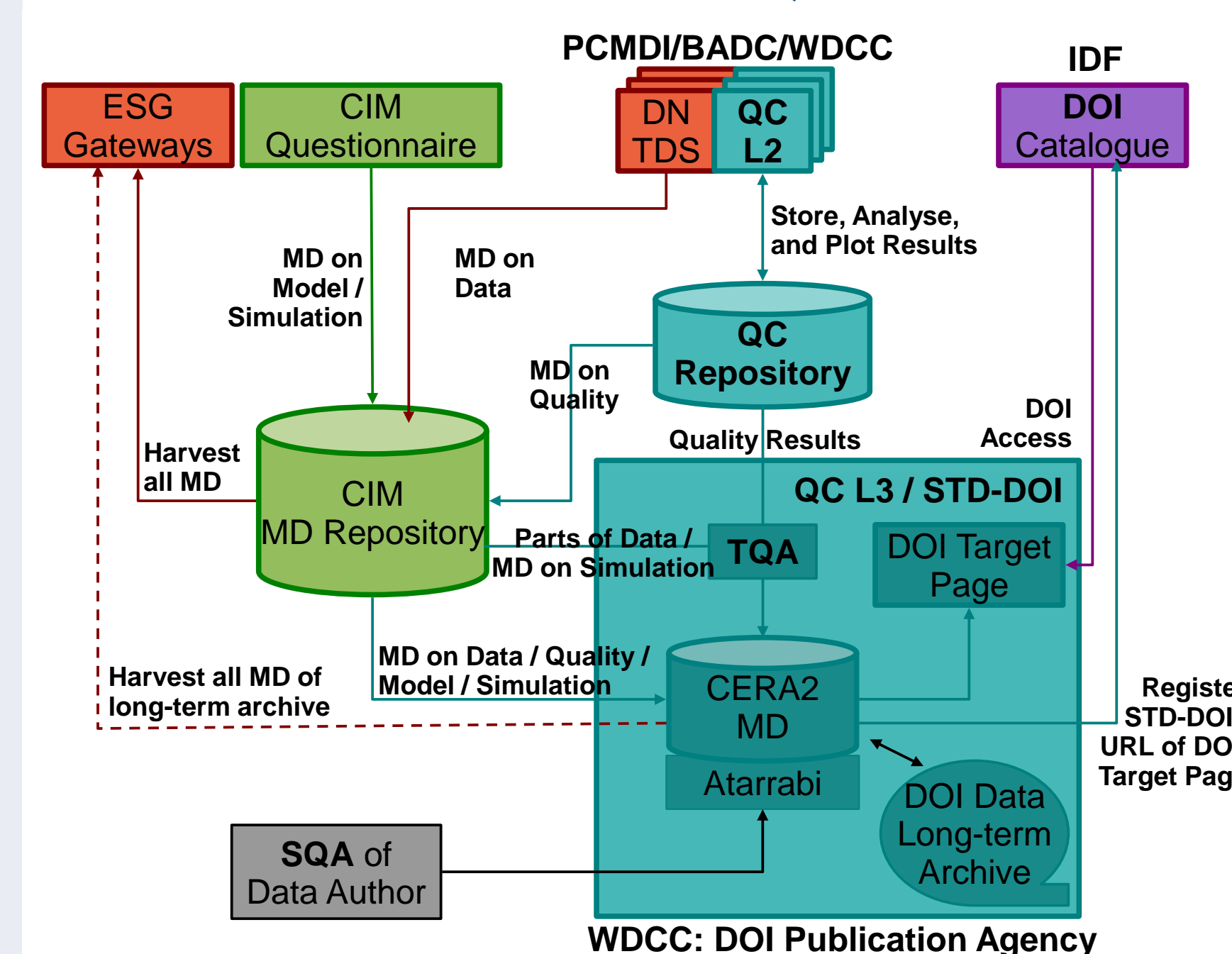


More Information:

<http://purl.org/org/cmip5/qc>

Distributed Quality Control Approach

Workflow of Distributed QC in CMIP5



ESG: Earth System Grid, MD: Metadata, DN: Data Node, TDS: Thredds Data Server, TQA: Technical Quality Assurance, SQA: Scientific Quality Assurance.

Distributed QC Procedure in CMIP5

CMIP5 data is delivered to one of the three ESGF partners, where it is ESG published and thus QC L1 Data checked. Afterwards QC L2 Data consistency checks are performed, before a data subset is replicated among the ESGF. QC L2 results are stored in a central QC Repository.

During QC L3 / DOI checks the QC results are accessed by the DOI Publication Agency WDCC. Other sources for cross- and double-checks are the CIM Metadata Repository, the Thredds Data Server (TDS), and the metadata stored in the long-term archive at WDCC.

Thus, the effort of the QC L2 Data checks is shared among the ESGF. But the QC L3 / DOI checks are performed at one site making use of the QC L2 results stored in a central QC Repository.

Distributed QC Approach

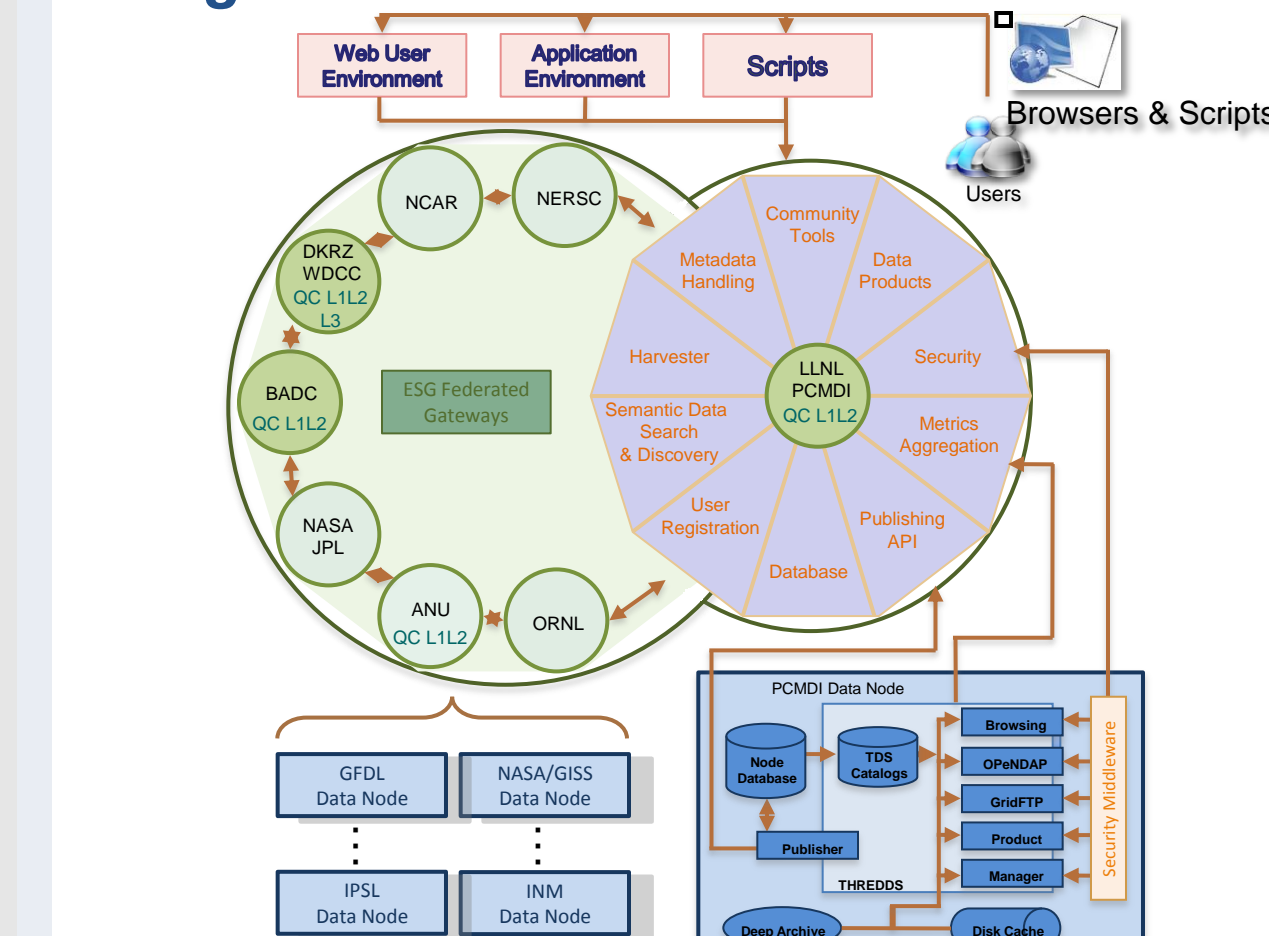
For high volume data such as climate data quality assurance has to be carried out at the data storage centre before opening the repository for data access. Data distributed in a Data Grid with its decentralized data repositories have to be checked at different sites with comparable QC procedures.

Thus a QC procedure/tools have to be developed, maintained and distributed centrally and agreed upon within the scientific community.

A distributed QC approach consists of different software components:

- **QC tool:** QC check
- **QC service layer (installed at QC L2 and L3 check sites):** QC result analysis, exception statistics, plotting, and QC Repository access
- **QC Repository:** central storage of QC results and check configuration

Organisation of CMIP5 Data Federation



More Information:

<http://purl.org/org/cmip5/qc>

Data Publication Procedure



Actors in DOI Publication Process

	Permission: QC L2	Scientific Q. Assurance	Technical Q. Assurance	DOI-Publication
Scientist				
Publication Agency				
Registration Agency				

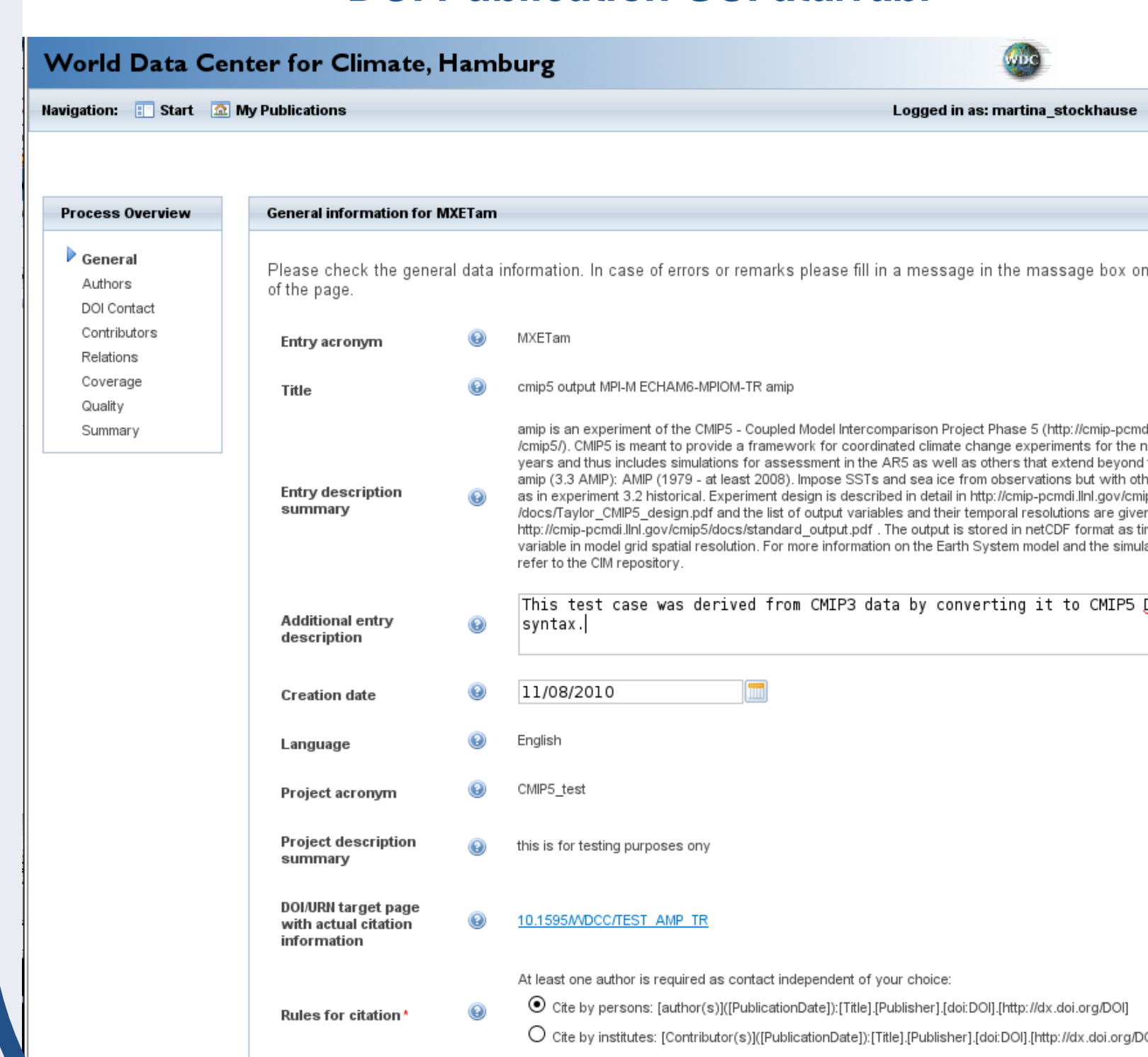
DOI Publication Process

The final DOI data publication procedure is in agreement with the regulations of the DataCite consortium:

- **Scientific Quality Assurance:** performed by the data author and documented via a publication service GUI (atarrabi)
- **Technical Quality Assurance:** cross- and double checks of data and metadata integrity
- **DOI Publication:** DataCite DOI metadata and DOI are separately send to the registration agency, a member of the DOI Foundation. Data and DOI remain unchanged and persistent.

DOI Construction Rule for CMIP5:
doi: 10.1594/WDCC/CMIP5.<opaque bit>

DOI Publication GUI atarrabi



More Information:

<http://cera-www.dkrz.de/atarrabi>



martina.stockhause@zmaw.de
WDC Climate / DKRZ: www.wdc-climate.de; www.dkrz.de
CMIP5: cmip-pcmdi.llnl.gov/cmip5; **CMIP5 Quality Control:** purl.org/org/cmip5/qc

