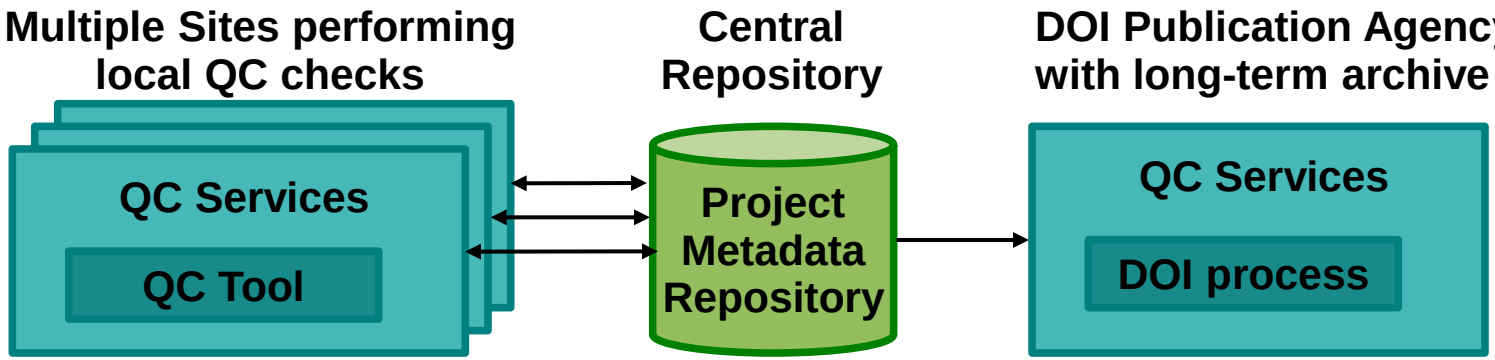
Quality Assessment Concept for CMIP5

Martina Stockhause, Michael Lautenschlager

Coupled Model Intercomparison Project (CMIP5)

- **Participants:**
  ca. 20 participating modelling centres
  with ca. 40 model configurations (different resolutions)
- **Experiments:**
  60 experiments with partly several realisations (ensemble members)
  over 90 000 model years
- **ca. 2 Mio. atomic datasets of ca. 400 experiments**
- **Data Volume:**
  - ca. 10 PB output: → **QC L1**
  - thereof ca. 2.5 PB requested ***'output1', 'output2'*** → **QC L2**
  - thereof ca. 1 PB replicated ***'output1':*** IPCC-AR5 → **QC L3 / DOI**

Distributed Quality Control Approach for High Data Volumes



QC checks:

- QC Execution Service (qcWrapper.py): QC tool run and Repository ingests of configuration and results
- QC Analysis Services (qcDbselect.py) for data analyses and exception statistics
- QC Plotting Service (qcDbselect.py) and plot ingest in Repository
- QC level assignment (qcAssignL2.py)

DOI publication:

- Export QC results (qcDbselect.py) for DOI publication process

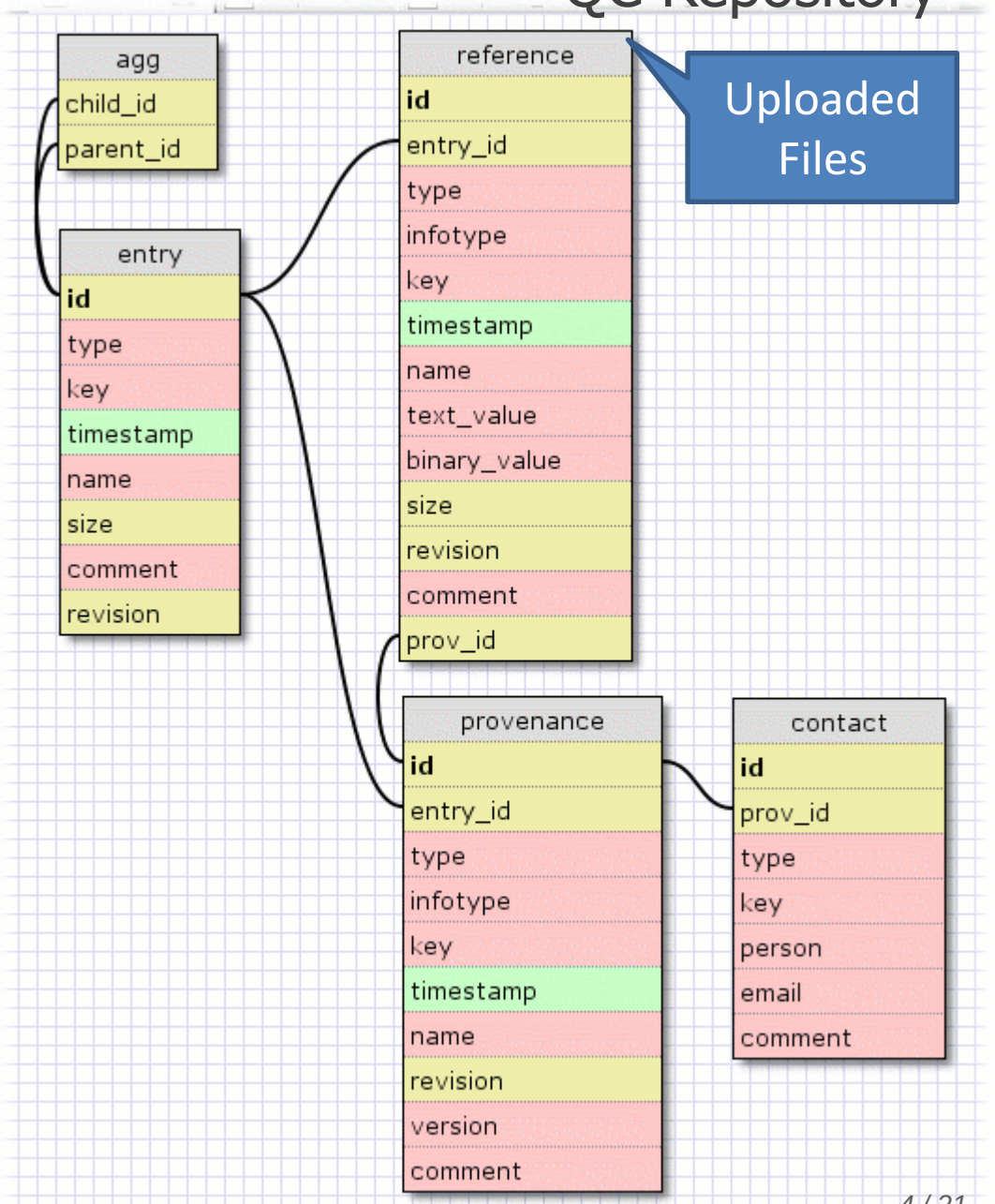
QC Repository

Aggregation Levels:

- **NC:** chunk names
  (for completeness control)
- **DRS Atomic Dataset:**
  QC results
- **DRS Experiment / DOI:**
  QC I/O files; plots; status; contact

Additional Services:

- View for version control
- Table with history information filled by a trigger



DKRZ
DEUTSCHES KLIMARECHENZENTRUM

Data Hierarchy Levels in QC

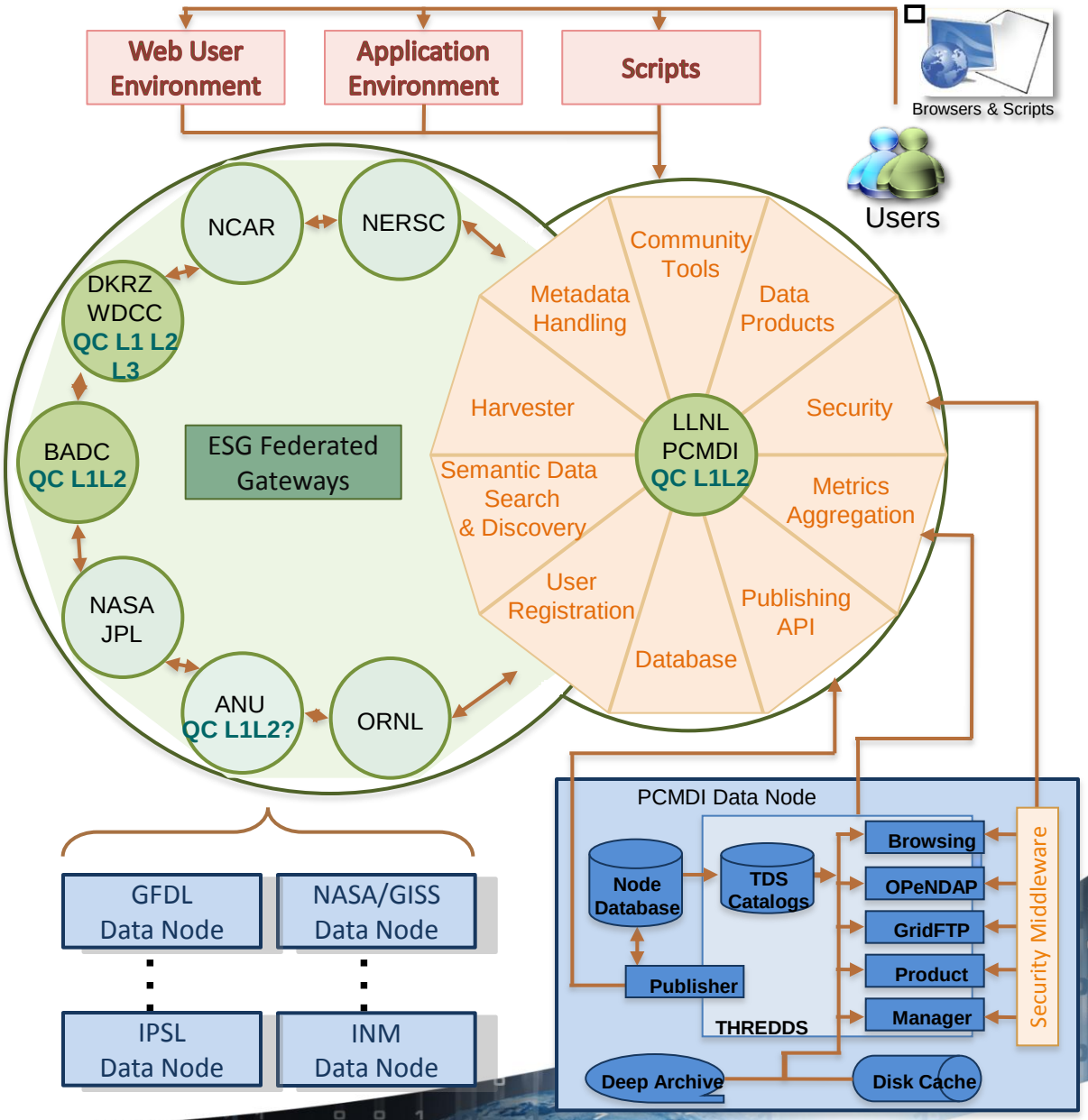
CMIP5 Process	DRS Name / Hierarchy Level
DOI / QC	<activity> <product> <institute> <model> <experiment>
	<frequency>
	<modeling realm>
	<MIP/CMOR2 table>
	<ensemble member>
Gateway Search	**<version>**
Atomic Dataset	<variable identifier>
Data Access	<netcdf datasets>

QC in CMIP5

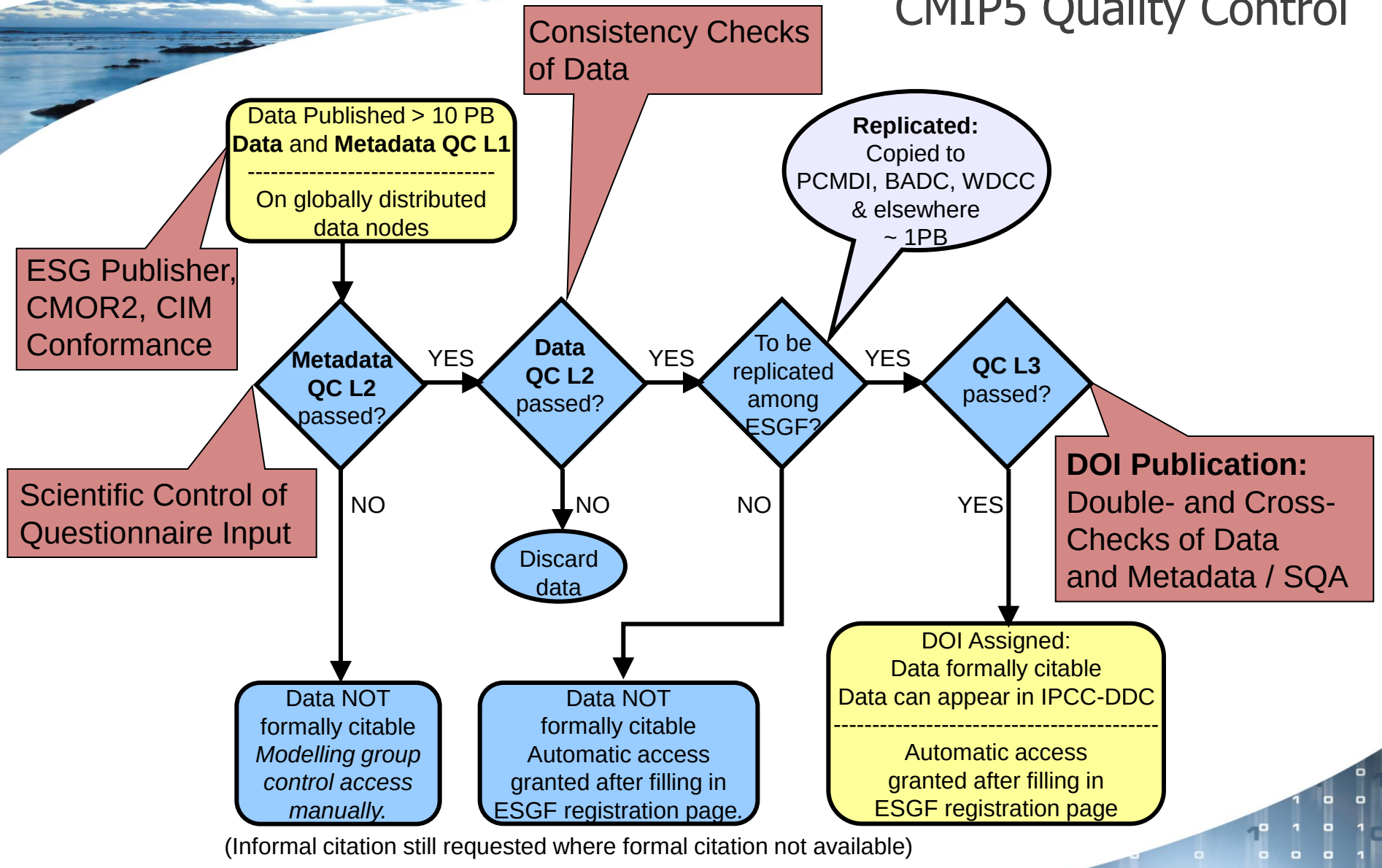
ESG Federation

PCMDI (Data / Security Earth System Grid: ESG)
BADC (Metadata: CIM / Metafor)
WDCC (Quality Control, DOI data publication)

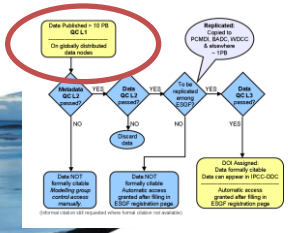
- Data Collection CMIP5
- ESG Gateway
- Data replication of CMIP5 data subset
- Quality Control (QC) on data for levels 1 and 2



CMIP5 Quality Control



(Informal citation still requested where formal citation not available)

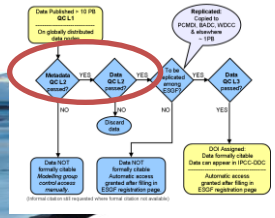


QC Level 1 (CMOR2 and ESG publisher conformance checks):

Performed at all ESGF partners during ESG publication

- **Data checks:**
  1. cmor2 compliance checks by the cmor checker check_CMOR_compliant.py
  2. esg publisher conformance

- **Metadata checks:** Completeness and technical validation of questionnaire input

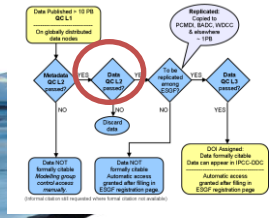


QC Level 2 (subjective quality control passed):

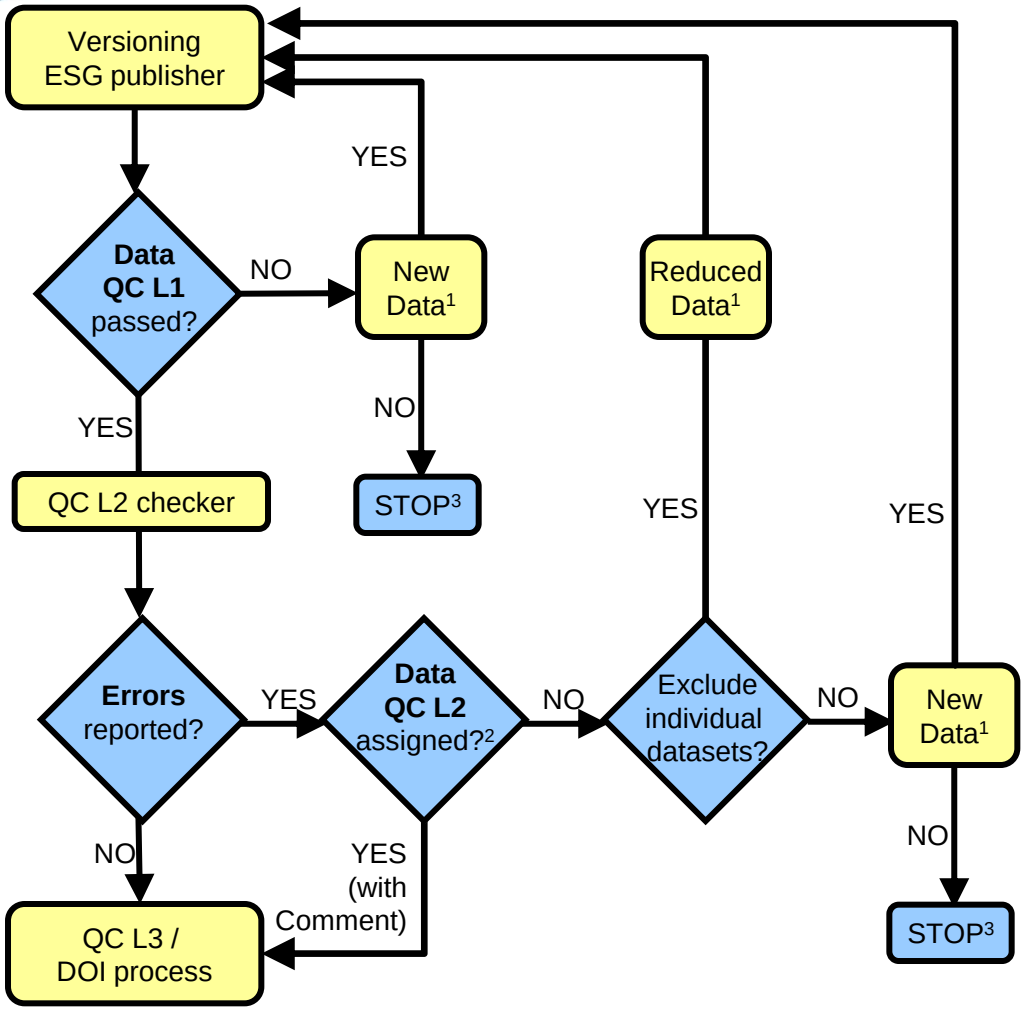
Performed on requested subset of CMIP5 data at all 3 (4) sites

- **Data checks:** Consistency checks: Check of statistical global values and additional DRS checks

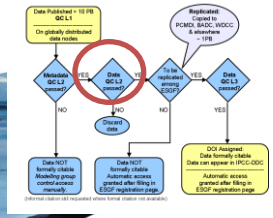
- **Metadata checks:** Subjective metadata control by scientist metadata available via AtomFeed



Assignment of Quality Control Level 2



[1] New version assigned to ESG datasets
[2] Assignment in co-operation with data author according to criteria http://www.leuchtturm-atlas.de/SCR/qc2list.html
[3] Delete possible QC L2 results out of QCDB



Assignment Criteria

Home Projects Help

CMIP5 » QC

QC2 Exceptions

http://www.leuchtturm-atlas.de/SCR/qc2list.html

QC2 Exception Codes (preliminary version)

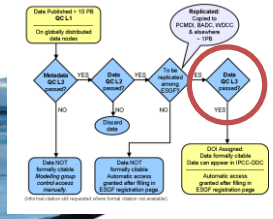
Fatals: immediate action neccessary,
errors: data unacceptable,
warnings: data possibly not ok,
informatory: fine - just for info,
unclear: title lines or open issues.

Thematic exception groups (see first column):
ACCEss errors, **GENE**ral checks, **META**data and file name checks,
TABL: inconsistencies in comparison to meta data tables,
TIMEaxis checks, **VARI**ables' checks.
OBSOlete messages (not used for CMIP5 project).

The following flags F<n> refer to general checks,
mainly on the time axis.

key group	description	comment
F-1 GENE	-- Not checked	
F0 GENE	-- No error found	
F1 TIME	testTimeStep() ^0, ^5 Error: negative time step	
F2 TIME	testTimeStep() ^0, ^5 Error: missing time step	This, of course, is no error if the QC is run over several time slices with intentional gaps in between. You may want to set NON_REGULAR_TIME_STEP in the setup file to check only for positive increments (of perhaps different sizes).
F4 TIME	testTimeStep() ^0, ^5 Error: identical time step	
F8 TIME	testCalendarTimeBounds() Error: negative/zero time bounds range	
F16 TIME	testCalendarTimeBounds() ^0 Error: overlapping time bounds ranges	
F32 TIME	testCalendarTimeBounds() ^0 Warning: gap between time bounds ranges	

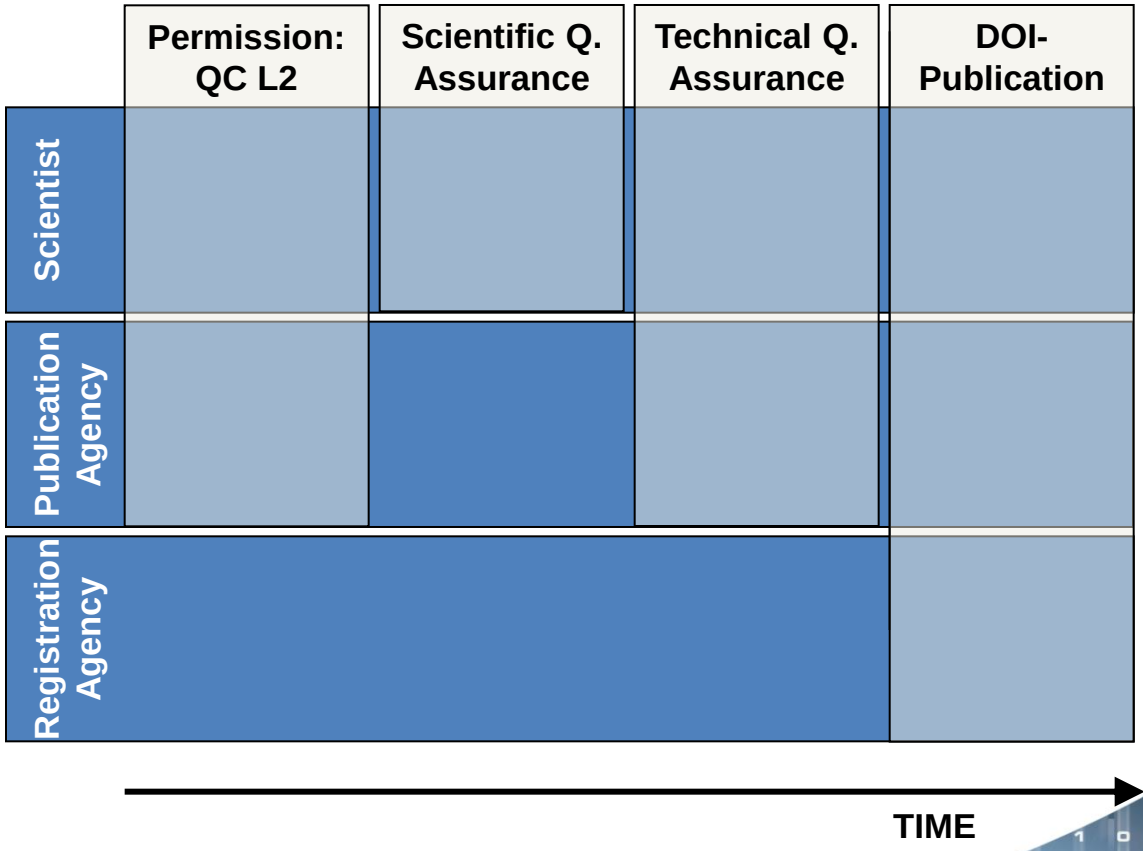
DKRZ
DEUTSCH
KLIMARECHENZENTRI



Quality Control Level 3 / DOI

QC Level 3 (approved by author):

- **Double and cross-checks of data and metadata**
- **Author check and approval of data and metadata (SQA)**
- **DOI publication of data (DataCite):** assignment of DOI as persistent identifier and citation direction resolving the DOI opens the DOI landing page.



E.g. doi:10.1594/WDCC/CMIP5.MXELr4

DOI landing

DOI for Scientific
10.1595/WDCC/TEST_AMP_TR

Title
cmip5 output MPI-M ECHAM6-M

Citation
Lautenschlager, Michael (2011)

Publication Date
2011-02-17

Contact for data entity
Joerg Wegner

CMIP5 Metadata hoste
http://cera-www.dkrz.de/WDCC/

Summary
amip is an experiment of the CI
experiments for the next five yea

3.3 amip (3.3 AMIP): AMIP (197

Experiment design is describe
pcmdi.llnl.gov/cmip5/docs/stan
The output is stored in netCDF
repository.

Quality
Accuracy: not filled

Consistency: Quality Control Le
* Level 0: Spot checks on selec
* Level 1: CMOR2 and ESG pub
* Level 2: Technical checks on t
* Level 3: Data approved by autl

Completeness: not filled

Specification: [CMIP5:QualityLe
[CMIP5:QualityControl2Comme
sub-T=6hrPlev, var=va, dim=tim
metadata not found or not acce

Data for CMIP5 experiment MXETam

Please choose your desired data destination. The displayed list of links refer to ESG datasets, i.e. collections of data belonging to a model realm / ensemble member (see DRS syntax definition [http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf]). Links resolve to the corresponding Thredds Data Server entries.

Location PCMDI

Dataset
cmip5 output MPI-M ECHAM6-MPIOM-TR amip mon atmos Amon r2i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip fx atmos fx r0i0p0 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip mon atmos Amon r1i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip mon landIce LImon r1i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip mon land Lmon r1i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip 6hr atmos 6hrPlev r2i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip day atmos day r1i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip day land day r1i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip day atmos day r2i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip mon land Lmon r2i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip 6hr atmos 6hrPlev r1i1p1 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip fx land fx r0i0p0 v20100928
cmip5 output MPI-M ECHAM6-MPIOM-TR amip mon landIce LImon r2i1p1 v20100928

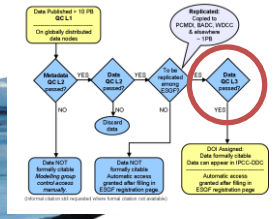
Back to the metadata page.

This page is hosted at WDCC, please send technical inquiries to data@dkrz.de.

Link to primary data
CMIP5Links.jsp?acronym=MXETam *(to be replaced by link into PCMDI gateway)*

Please note
WDC Climate as DOI publishing agency grants the validity of data accessed at WDC Climate. It recommends to check all downloaded data files by their tracking_id with the CMIP5 Data Validation Service at http://cera-www.dkrz.de/CMIP5Tracking.jsp?tracking_id=*tracking_id*. The tracking_id can be found in the file headers (e.g. using ncdump -h).

This page is hosted at WDCC/DKRZ, please send technical inquiries to data@dkrz.de.



GUI for Scientific Quality Assurance

Atarrabi:

- GUI for final author check of selected metadata (general, contact, authors, contributors, citations,…)
- Scientific quality information should be added
- Publication of environ. data: umwelt.wikidora.com

cera-www.dkrz.de/atarrabi/

Process Overview

- General
- Authors
- DOI Contact
- Contributors
- Relations
- Coverage
- Quality
- Summary

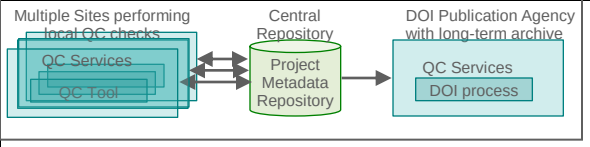
General information for MXETam

These general data information are shown on the DOI/URN target page and library catalogues like GetInfo. The core metadata properties are chosen for accurate and consistent data identification in citations and for data retrieval, along with recommended use instructions. They belong to the metadata kernel of the DOI (STD-DOI).

Entry acronym	MXETam
Entry name	cmip5 output MPI-M ECHAM6-MPIOM-TR amip
Entry title *	cmip5 output MPI-M ECHAM6-MPIOM-TR amip
Entry description summary	amip is an experiment of the CMIP5 - Coupled Model Intercomparison Project Phase 5 (http://cmip-pcmdi.llnl.gov /cmip5/). CMIP5 is meant to provide a framework for coordinated climate change experiments for the next five years and thus includes simulations for assessment in the AR5 as well as others that extend beyond the AR5. 3.3 amip (3.3 AMIP): AMIP (1979 - at least 2008). Impose SSTs and sea ice from observations but with other conditions as in experiment 3.2 historical. Experiment design is described in detail in http://cmip-pcmdi.llnl.gov/cmip5 /docs/Taylor_CMIP5_design.pdf and the list of output variables and their temporal resolutions are given in http://cmip-pcmdi.llnl.gov/cmip5/docs/standard_output.pdf . The output is stored in netCDF format as time series per variable in model grid spatial resolution. For more information on the Earth System model and the simulation please refer to the CIM repository.
Additional entry description	
Creation date	11/08/2010
Language	English
Project name	IPCC/CMIP5_test
Project description summary	this is for testing purposes ony
DOI/URN target page with actual citation information	http://cera-www.dkrz.de/WDCC/ui/CMIP5Compact.jsp?acronym=MXETam
Rules for citation *	At least one author is required as contact independent of your choice: (•) Cite by persons: [author(s)]([PublicationDate]):[Title].[Publisher].[doi:DOI].[http://dx.doi.org/DOI] ( ) Cite by institutes: [Contributor(s)]([PublicationDate]):[Title].[Publisher].[doi:DOI].[http://dx.doi.org/DOI]
CMIP5 questionnaire *	☑ I approve that the metadata I filled in the CMIP5 questionnaire and sent to CMIP5 is correct and up-to-date.

Message to publication agent

DKRZ DEUTSCHES KLIMARECHENZENTRUM



CMIP5 Distributed Quality Control Design

PCMDI/BADC/WDCC

IDF

ESG Gateways

CIM Questionnaire

DN TDS

QC L2

DOI Catalogue

Store, Analyse, and Plot Results

MD on Model / Simulation

MD on Data

Project Metadata Repository

QC Repository

MD on Quality

Quality Results

DOI Access

Harvest all MD

CIM MD Repository

QC L3 / STD-DOI

Parts of Data / MD on Simulation

TQA

DOI Target Page

MD on Data / Quality / Model / Simulation

Harvest all MD of long-term archive

CERA2 MD

atarrabi GUI

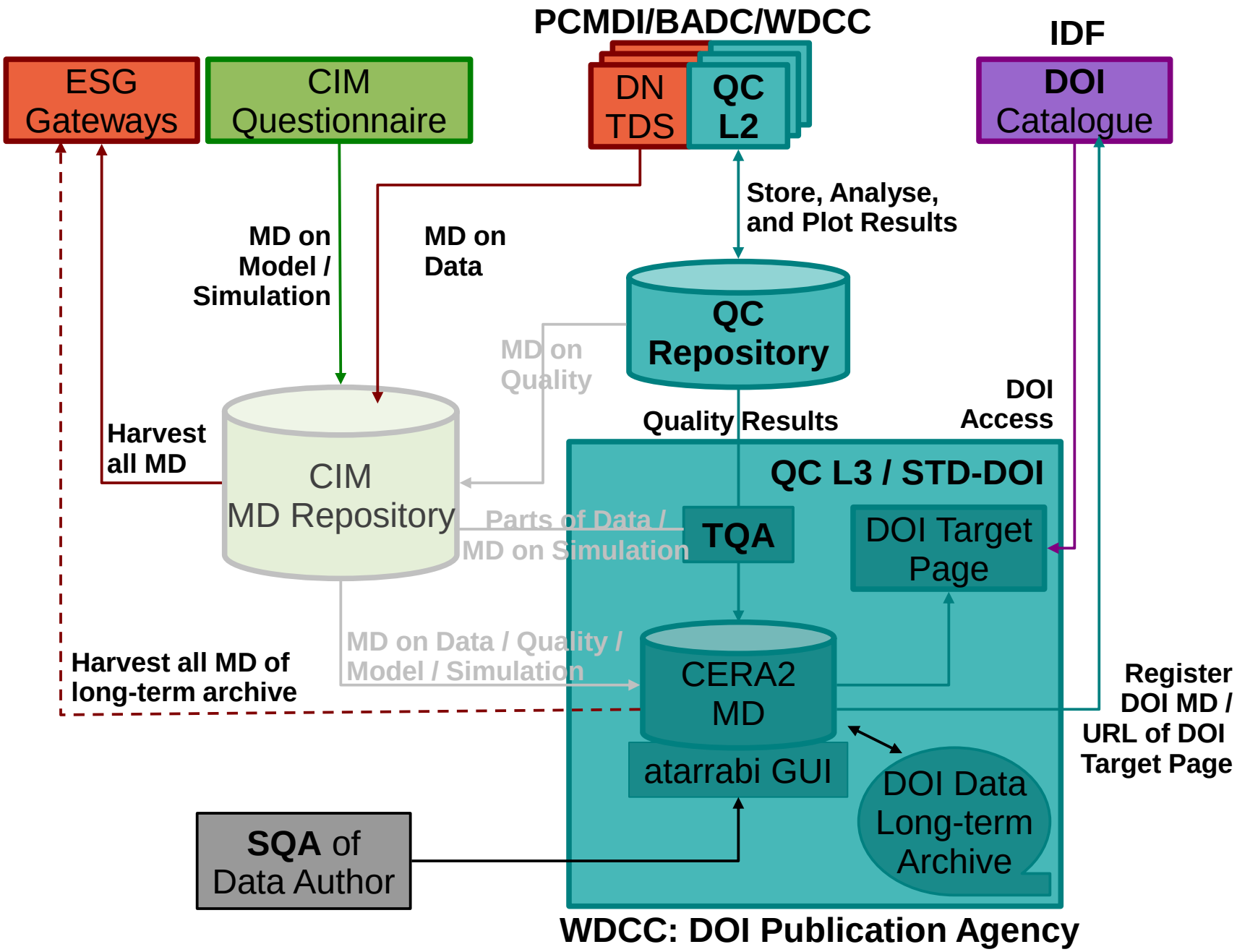
Register DOI MD / URL of DOI Target Page

DOI Data Long-term Archive

SQA of Data Author

WDCC: DOI Publication Agency

CMIP5 Distributed Quality Control Status



QC Development:

Status QC in CMIP5 (1)

- **QC L2:**
  QCDB moved from postgres to oracle for production
  due to maintenance reasons (21st April 2011);
  bug-fixes in QC tool
- **QC L3:**
  DOI data publication moved from STD-DOI to DataCite (21st April 2011)
  SQA GUI atarrabi final release 1.5 (2nd May 2011):
  cera-www.dkrz.de/atarrabi/
- **DOI landing page:**
  set up: cera-www.dkrz.de/CMIP5/CMIP5Compact.jsp?acronym=<acronym>
  design under discussion:
  cera-www.dkrz.de/CMIP5/CMIP5example.html
  data set access realized by a list of ESG dataset TDS entry points because of the DOI granularity on DRS experiment level
- **DOI services:**
  - tracking_id service: cera-www.dkrz.de/CMIP5/CMIP5Tracking.jsp
  - view for QC status/DOI in place: cera2.v_qc_status
  - view for ESG datasets belonging to DOI in place: cera2.v_ipcc_files

CIM access: Status QC in CMIP5 (2)

- **Get Contact/Authors/Title from simulationRun object:**
  no access of AtomFeed entries by DRS experiment name
  -> use of persons from email receiver list for QC L3 process
- **Ingest of QC L2:**
  QC questionnaire for QC check definition exists
  Ingest tool for QC result ingest into CIM missing
  no QC flag communication to gateways, no QC results in CIM
  -> use WDCC view cera2.v_qc_status (cera2.v_ipcc_files)
- **Update CIM after QC L3 / DOI publication:**
  unclear
  no QC flag/DOI communication to gateways, no update in CIM
  -> use WDCC view cera2.v_qc_status (cera2.v_ipcc_files)
- **DOI landing page:**
  no link to CIM metadata available so far
  -> use link to CERA metadata in the meantime

Technical QC Meeting planned in UK in May 24 + 25th, 2011

Open Issues (1)

1. **Will ANU perform QC L2 checks on the Australian data?**

a. Yes, Australian data. Contact person needed.

b. No.

2. **How will the gateway show QC flag and DOI?**

a. For the DOI / DRS experiment and the ESG datasets:
Then we could use a single link out of the DOI landing page into the ESG gateway. Harvest the information from:

I. CIM ?
II. WDCC: Views cera2.v_qc_status und cera2.ipcc_files

b. Only for DOI / DRS experiment:
Then we do not have a connection between DOI and the datasets belonging to it in the gateways. Harvesting:
I. CIM ?
II. WDCC: View cera2.v_qc_status

Open Issues (2)

3. How do we deal with missing or unfinished CIM metadata descriptions?

a. Part of CIM metadata might be sufficient but even not ideal: Specification of mandatory parts or percentage of completeness of CIM needed.

I. QC on data is assigned QC L2 without CIM metadata. For DOI publication CIM metadata have to exist, i.e. published by AtomFeed.
II. QC stops before assignment of QC L2 if no or insufficient CIM metadata exists.

b. CIM metadata is regarded as optional: QC L3 is finished by DOI assignment only with TDS and DOI metadata. Data Authors have to complete metadata for DOI data publication.

4. How do we ensure identical QC L2 data checker application at all three (four) sites? This is crucial for QC-L3 and the DOI publication process.

a. Use of identical software version of QC tool and in same configuration.

b. Otherwise during QC L3 the QC L2 checks have to be inferred and completed if necessary.

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.

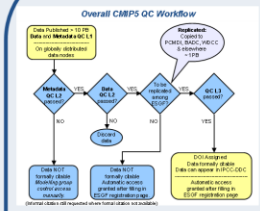
Because of the high data volume the QC checks up to level 2 are performed distributed among the ESGF. The final QC level 3 checks for the DOI assignment are carried out by WDCC. Afterwards CMIP5 data is formally citable and remains persistent.

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 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).

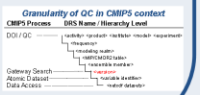
Quality Levels for CMIP5

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

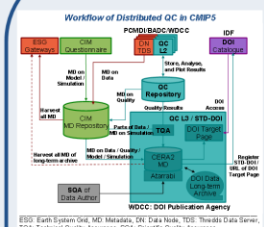
QC checks for data and metadata are performed, separately, for levels 1 and 2. During the cross-checks of QC L3 checks their results are reviewed.

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



More Information: http://purl.org/org/cmip5/qc

Distributed Quality Control Approach



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

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.

Our *distributed QC approach* consists of different software components:

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.

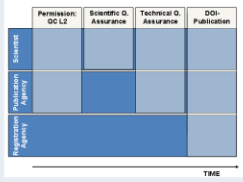
Thus, the effort of the QC L2 Data checks is shared among the ESGF. But the QC L3 / DOI checks are performed at one site making use of the QC L2 results stored in a central QC Repository.

Organisation of CMIP5 Data Federation

More Information: http://purl.org/org/cmip5/qc

Data Publication Procedure

Actors in DOI Publication Process



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

QC L3 / DOI Process in CMIP5

The cross- and double checks of the Technical Quality Assurance make use of the QC result of the preceding levels. Data as well as metadata is reviewed and data accessibility checked.

The final author approval step is supported by the GUI atarrabi. Authors check basic metadata and add information about their own quality assurance (Scientific QA).

A DOI is assigned and registered at the International DOI Foundation (IDF: dx.doi.org) via the Registration Agency DataCite.

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

http://cmip-pcmdi.llnl.gov/cmip5/

http://cmip5qc.wdc-climate.de