

Long-term archiving workflow in CMIP5

- a first review -

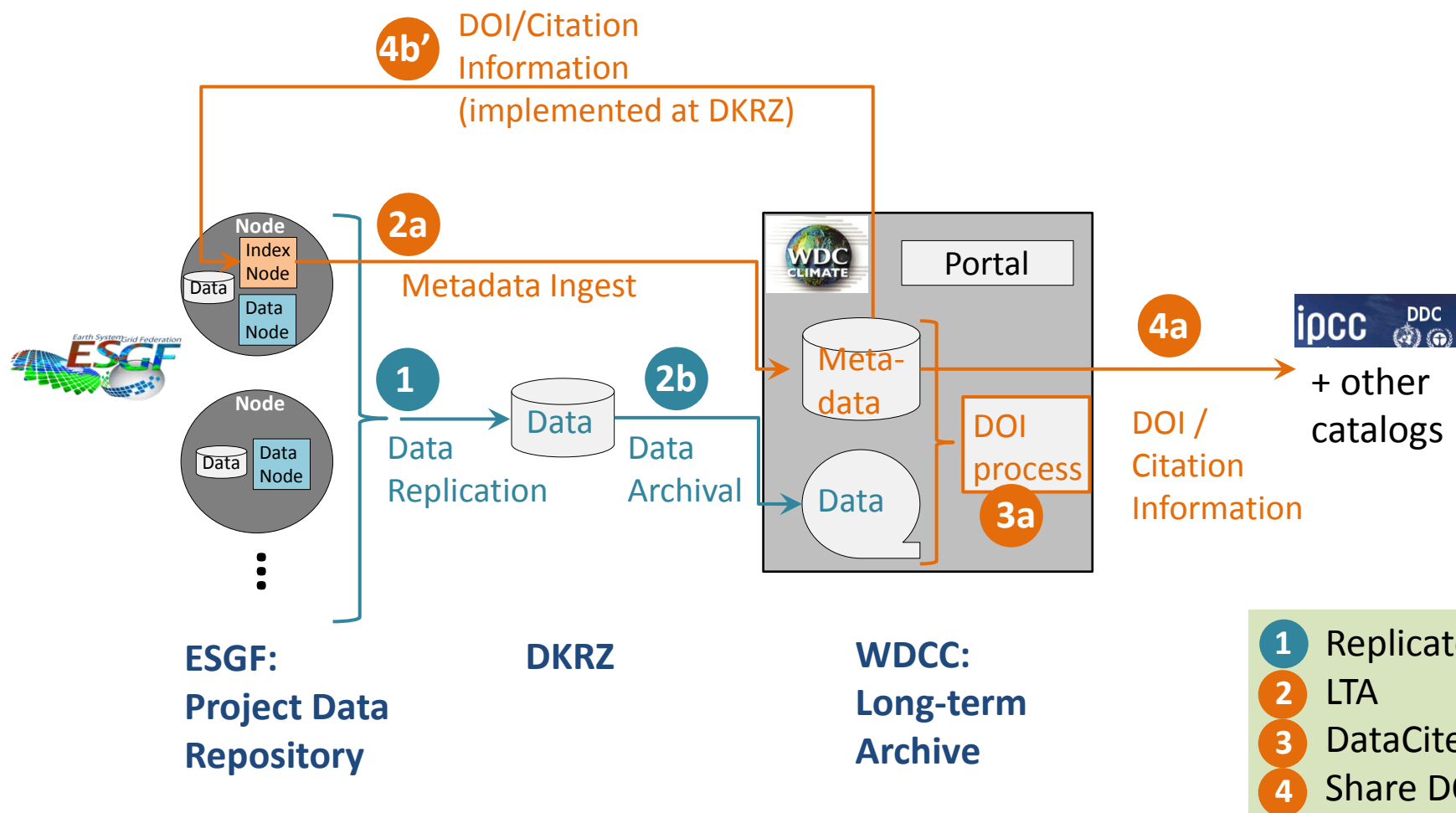
Martina Stockhause
Deutsches Klimarechenzentrum (DKRZ)

Acknowledgement: Most colleagues of the Data Management department of DKRZ were involved in the execution of the CMIP5 LTA workflow.

Long-Term Archival in CMIP5

- The purpose of Long-term archival (LTA) and the IPCC DDC is to provide **stable data for long-term interdisciplinary (re-)use**:
 - Permanent and persistent access to
 - stable data
 - of high-quality and
 - well-documented.
- LTA and the following quality assurance process connected with the DataCite DOI process and the integration into the IPCC DDC AR5 are **no integral parts of the CMIP5 data infrastructure**.

Simple LTA Workflow



- 1** Replicate
- 2** LTA
- 3** DataCite DOI
- 4** Share DOI

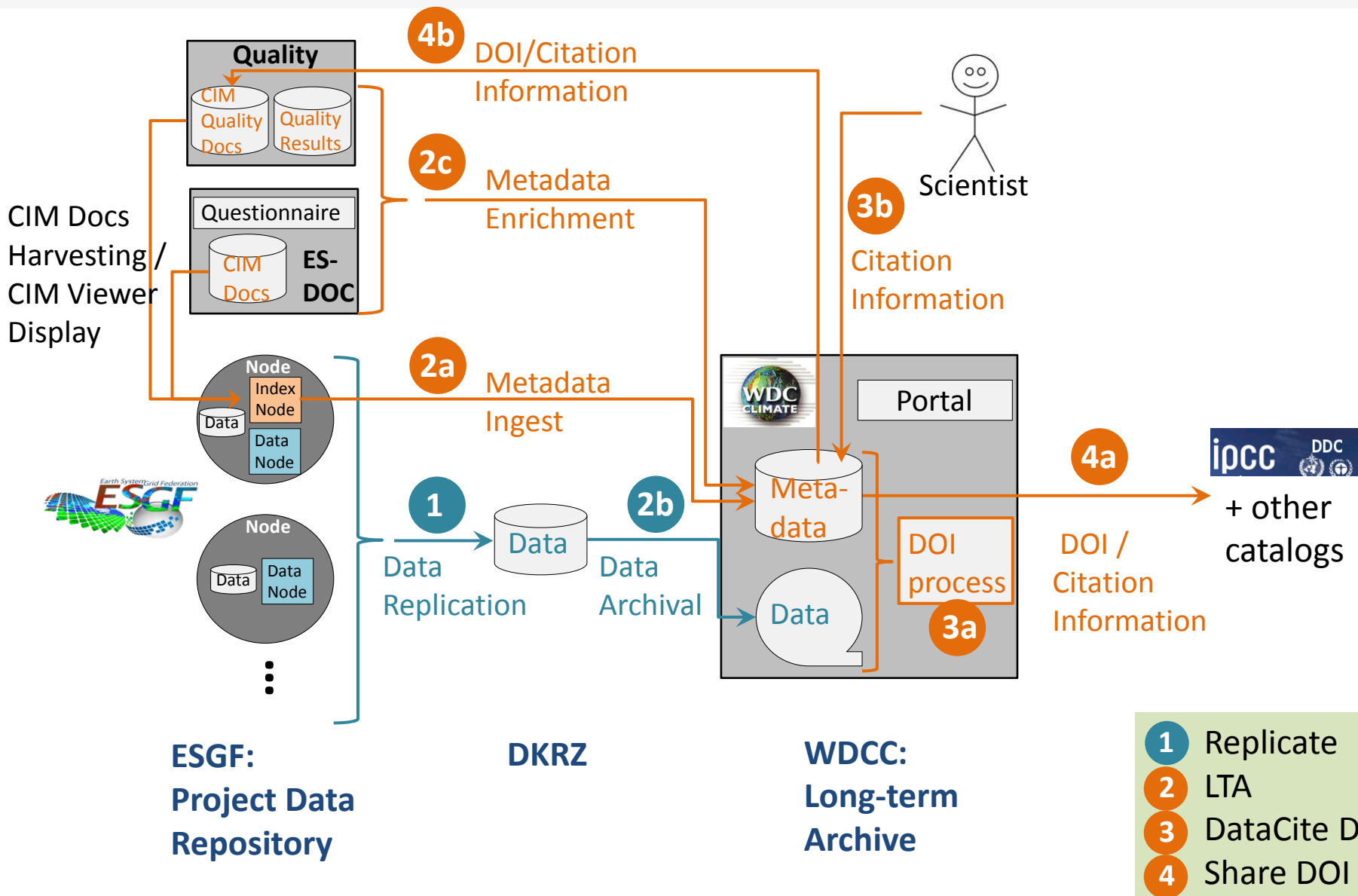
Identification of Data Objects (1)

Identifier

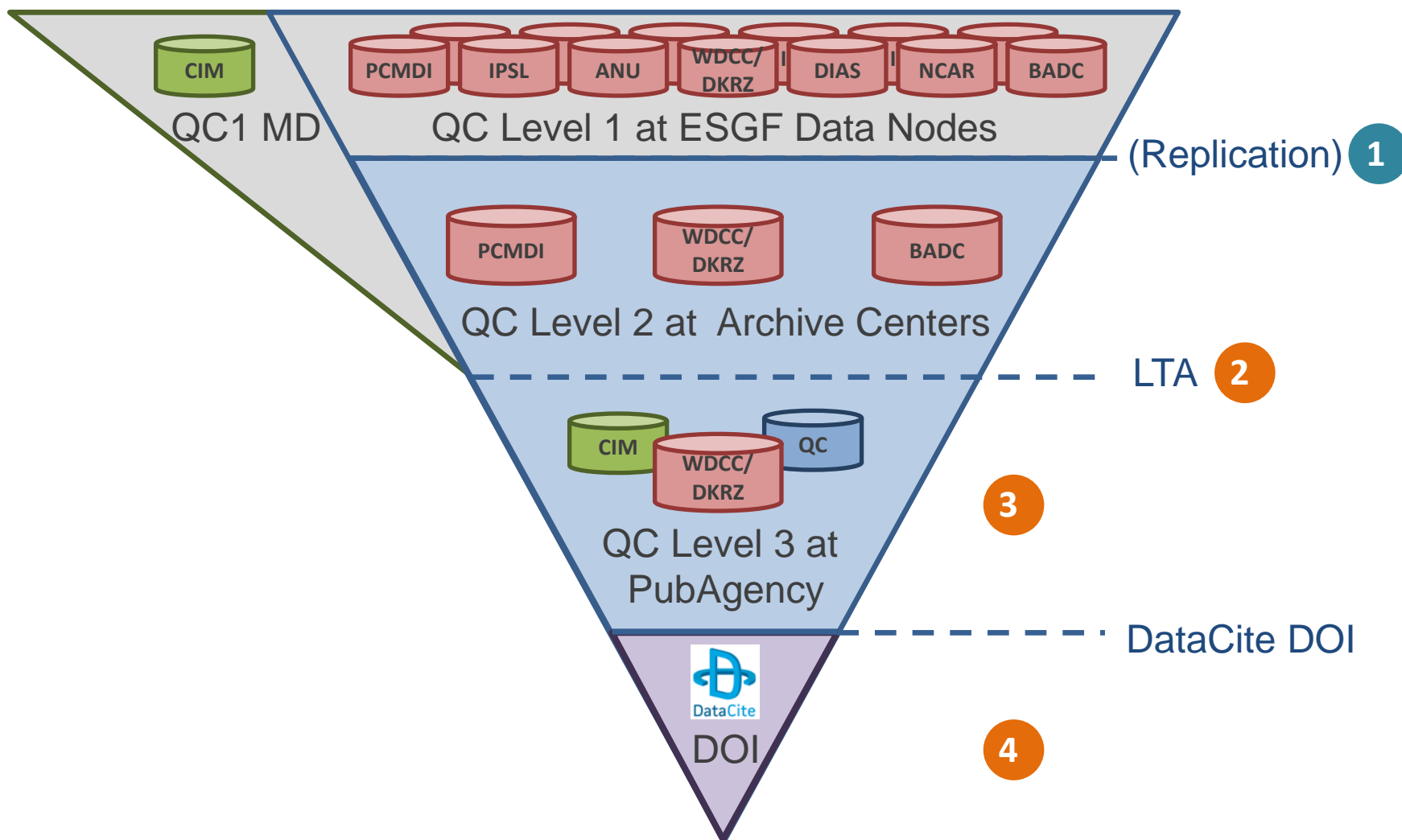
ID	Unique?	data header	ESGF portal
tracking_id	(yes)*	yes	yes
DRS_id + version	<ul style="list-style-type: none"> • 2 types of DRS_ids • Model/institute names in 2 variations • Same file published in different versions 	DRS_id without version (type 1)	DRS_id with version (type 2)
checksum	(yes)*	no (can be calculated)	yes

* Unique by design but correct usage not enforced
(Unique if created with designated software
but not unique if not recreated for changed data.)

LTA Workflow with CIM and QC...



CMIP5 Quality Control



Identification of Data Objects (2)

Identifier

ID	Unique?	data header	ESGF portal	CIM / ES-DOC	Quality
tracking_id ⁺	(yes)	yes	yes	--	yes
DRS_id with version ⁺	<ul style="list-style-type: none"> • 2 types of DRS_ids • Model/institute names in 2 variations • Same file published in different versions 	DRS_id without version (type 1)	DRS_id with version (type 2)	no (uses own DRS-like externalIDs)	DRS_id with version (type 2)
checksum ⁺	(yes)	no (can be calculated)	yes	--	no
CIM uid	yes used by CIM	--	no	yes	--
File size ⁺	no additionally used within Quality	yes	yes	--	yes

⁺ cross-checked during DOI quality assurance process **3**

Identification of Data Objects (3)

Controlled Vocabularies:

- DRS component names:
 - Two slightly different versions are allowed for display (search facets) and in the file system, e.g. BCC-CSM1.1(m) vs. bcc-csm1-1-m
 - Some documented institute and model names on CMIP5 project pages differ from those in ESGF
 - Some experiment, institute, and model names differ between ESGF and ES-DOC (mostly corrected in ES-DOC)
 - Two types of DRS for production and for the ESGF: Differences between ESGF and data headers in version, product, ensemble member (rip vs. realization numbers)
- CF standard names: not checked against CV but against Taylor's standard output definitions
- ES-DOC CVs: used as LOVs in the questionnaire

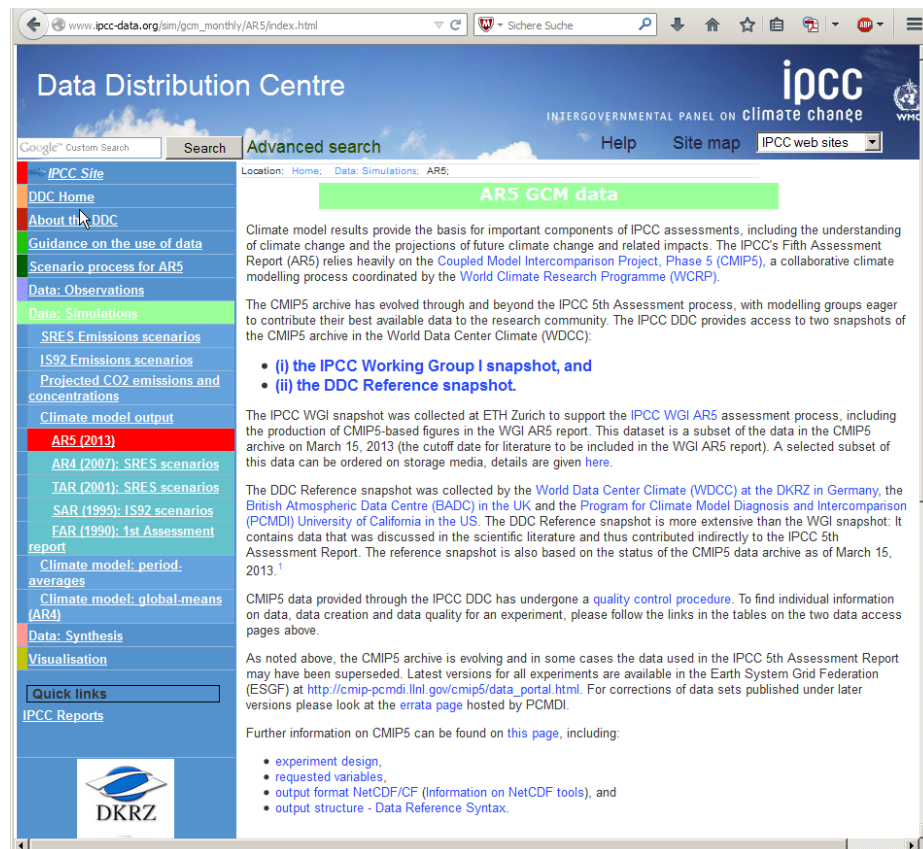
Interfaces in the LTA Workflow

- ESGF Search API
 - 1 Manage the data replication process
 - 2a Generate use metadata for long-term archive (LTA)
 - 3a Validation of data replica against original data
- Atom Feed (CIM)
 - 2c Access CIM simulation document for QC3 and enrichment of LTA metadata, e.g. experiment description
 - 4b Provision of QC results and formal citation by DataCite DOIs
- DataCite API
 - 4a Register DataCite citation metadata and mint DOI
- Other:
 - 2c Python based database access of Quality results to enrich LTA metadata (QCWrapper), DB views for QC3 cross-checks
 - 3b GUI atarrabi to support the DOI process and to collect citation information from modeling groups

IPCC-DDC AR5 – Status of the CMIP5 LTA WF

IPCC DDC AR5 consists of two data collections (snapshots 15.03.2013):

- IPCC Working Group I snapshot (collected by the ETH Zurich): *complete*
- DDC Reference snapshot (transferred from ESGF into LTA): *LTA not yet finished*



The screenshot shows the IPCC Data Distribution Centre (DDC) website. The main heading is "Data Distribution Centre" with the IPCC logo and "INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE". The page is titled "AR5 GCM data" in a green banner. Below this, there is a section titled "Climate model results provide the basis for important components of IPCC assessments..." followed by a paragraph explaining the IPCC Working Group I snapshot and the DDC Reference snapshot. A list of bullet points includes:

- (i) the IPCC Working Group I snapshot, and
- (ii) the DDC Reference snapshot.

 Further text describes the IPCC WGI snapshot collection at ETH Zurich and the DDC Reference snapshot collection at the World Data Center Climate (WDCC) at DKRZ in Germany, the British Atmospheric Data Centre (BADC) in the UK, and the Program for Climate Model Diagnosis and Intercomparison (PCMDI) University of California in the US. A "Quick links" section is visible at the bottom left of the page content, and a DKRZ logo is at the bottom center.

http://ipcc-data.org/sim/gcm_monthly/AR5/

Review (1): What went *not* so well...

- Redesign of data infrastructure from ESG/trackback to ESGF/CIM Viewer:
 - Replication changed (use of ESGF Search API)
 - Metadata harvesting changed (data nodes to index node)
 - ESGF provides no URLs for use in external references (DOI landing pages)
 - Redesign of CIM document integration in ESGF portals required which has still errors in the DOI information (integration of Quality/Citation information)
- > all LTA interfaces to ESGF changed
- Policies on versioning and the use of identifiers were unclear or not strictly followed
- > Data Inconsistencies within ESGF affecting data replication
- No deadline for IPCC-DDC AR5 data delivery and no core data subset but dynamic CMIP5 data
- > Data Replication was finished (too) late; delay in LTA/DOI
- Lack of Controlled Vocabulary (DRS_id)
 - No central repository for controlled vocabulary to check against
 - late integration in CIM with variations after end of Metafor by ES-DOC
- > Mapping of model and institute names required

Review (2): What went well...

- Collaboration within ESGF and ES-DOC
- Data versioning allows to identify revised data
- Distribution of a large amount of data shortly after data creation by ESGF (decentralized data publication)
- Availability of detailed background information for the data (CIM documents)
- Reliable formal data citation by DataCite DOIs
- Quality checks added value to the data

What could be improved for CMIP6 (1)

■ Project administration:

- *Joint infrastructure development* of CMOR2, ES-DOC and ESGF with stable technical interfaces and clear timelines
- Development of clear *policies* for data quality, versioning etc.
- Central repository for *controlled vocabulary (CV)*, e.g. model and institute names
- Definition of *core* data (selected experiments and variables for the DDC)
- Improved *interaction with data creators*: Central entry point for modeling centers to enter information on CV, simulations, data volume, citation information, errata, annotations etc.

What could be improved for CMIP6 (2)

■ CMOR2:

- Provide identifiers in netCDF headers with links or PIDs to external information, e.g. use `tracking_id` as PID during ESGF data publication or provide links to simulation description (ES-DOC) / used CV...

■ ESGF:

- *Enforcement* of consistent use of identifiers and data versioning and other agreed *policies*, e.g. by use of PIDs
- Provision of *dataset URLs* within ESGF to point to them externally; for data citation a possibility for the *verification of specific data collections* is needed (e.g. an experiment, which were latest versions at a certain time in the past)
- *Integration of additional metadata* into ESGF, e.g. searchable selected CIM/Quality/Citation/Annotation/Provenance metadata

What could be improved for CMIP6 (3)

■ Citation:

- **Informal:** collect data citation information with the data (CMOR2?), ideally with PID assignment
- **Formal** (integration in reference lists of scientific papers) requires stable data, long-term archived (LTA) at a reliable data center.

Investigation for CMIP6 of:

- Speed-up possibilities for the data DOI process
- Decentral LTA and DOI processes by several DataCite DOI publication agencies
- Develop clear citation rules for dynamic data (within RDA/WDS IG/WGs on Publishing Data), e.g. pre-print DOIs (verification criteria)

■ External Data Services?

What could be improved for CMIP6 (4)

*Closely data-related information entered via **ESGF**:*

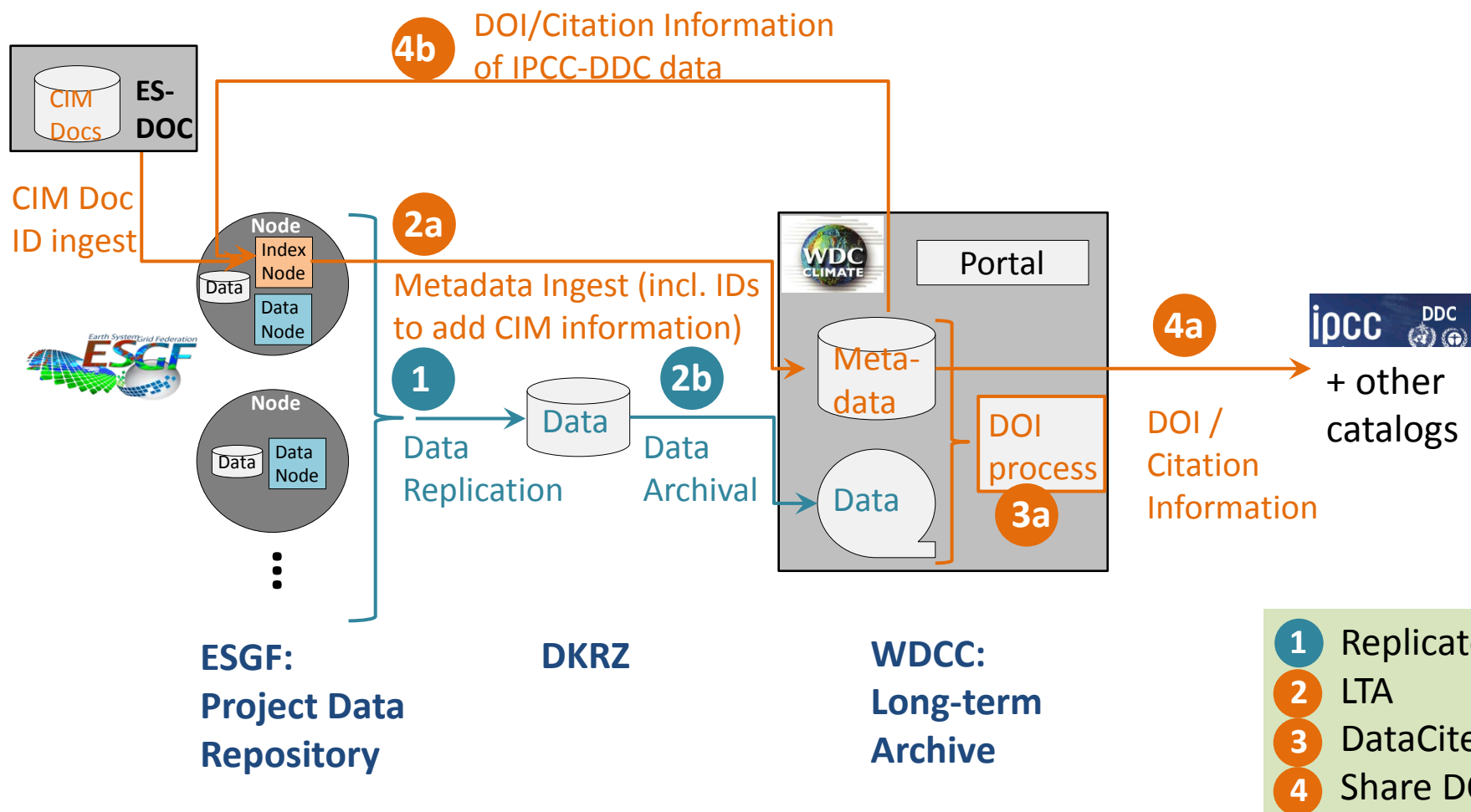
- Use metadata
- Data quality
- Data citation
- Data provenance
- Data annotations

*Information for data interpretation / reuse entered via **ES-DOC**:*

- Experiment
- Model
- Simulation
- Platform

ESGF and ES-DOC portals provide two entry points to CMIP6 information, which need to be well-connected

Ideal LTA Workflow



CMIP5: <http://cmip-pcmdi.llnl.gov/cmip5/>

IPCC-DDC AR5: http://ipcc-data.org/sim/gcm_monthly/AR5/

ESGF: <http://esgf.org/>

ES-DOC: <http://es-doc.org/>

CMIP5 QC: <http://cmip5qc.wdc-climate.de/>

DataCite: <http://datacite.org>

Stockhouse et al. (2012): Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data, Geosci. Model Dev., 5, 1023–1032, [doi:10.5194/gmd-5-1023-2012](https://doi.org/10.5194/gmd-5-1023-2012)