Long-term archiving workflow for CMIP6

ISENES2 Workshop on ESM Workflows 28.09.2016

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.

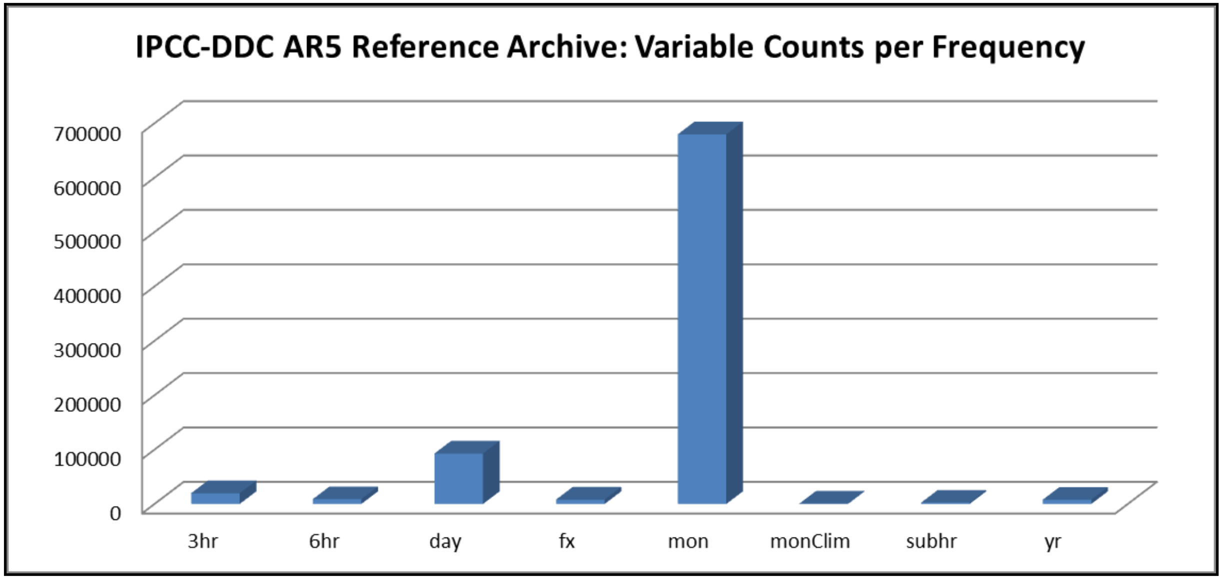
I. Long-Term Archival (LTA) in CMIP5

(IS-ENES Workshop on Workflows 04.06.2014: doi:10.5281/ZENODO.29104)

Looking Back: Long-Term Archival for CMIP5 (1)

The DDC Reference Archive / The IPCC WG1 Archive

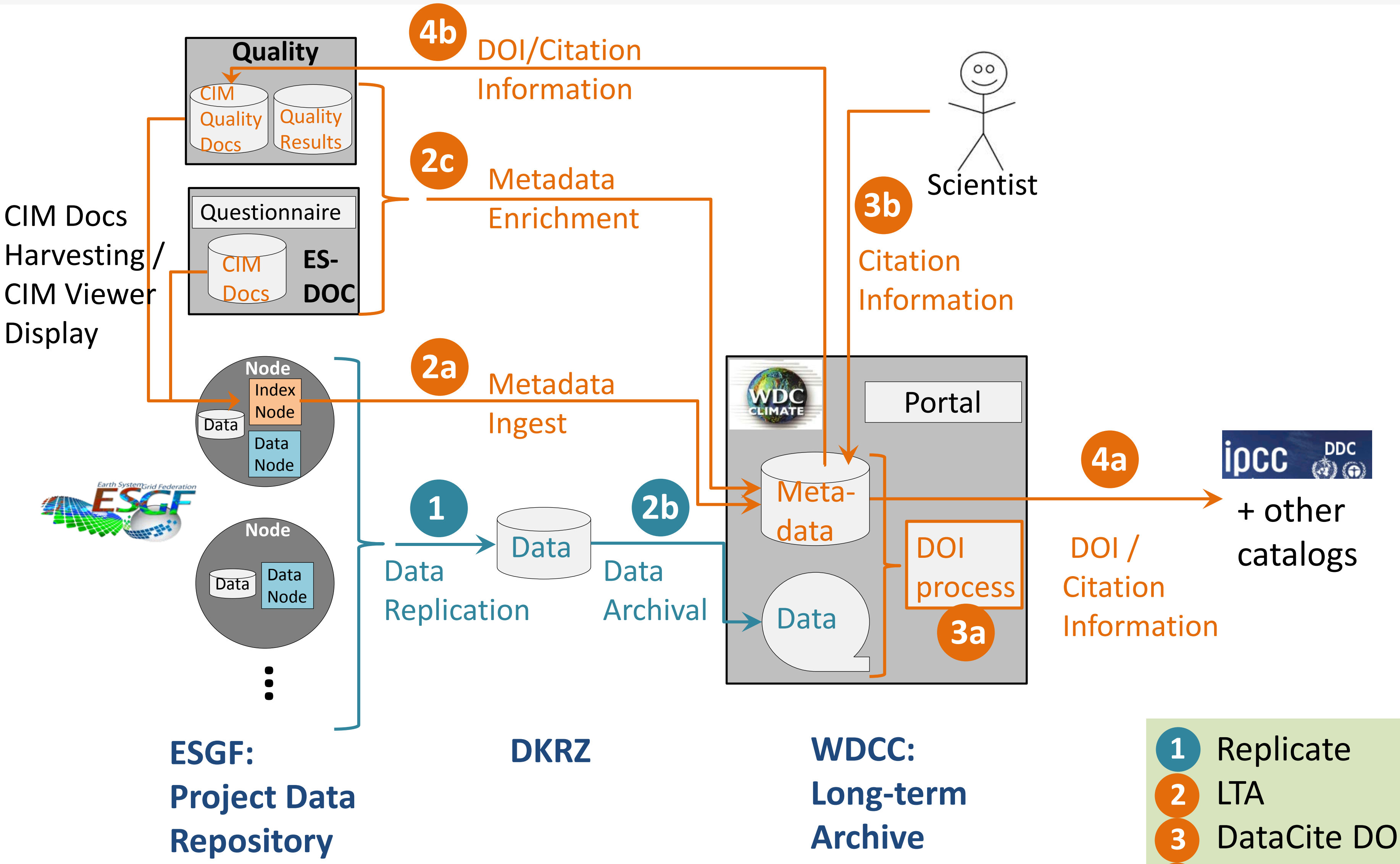
Experiments:	101 / 78	different experiments / scenarios
Variables:	605 / 123	different variables (628 requested variables)
Size:	1.6 PByte / 100 TByte	(all AR data: 1.7 PByte)
Models:	60 / 58	participating models
Institutes:	27 / 24	participating institutes
Simulations:	1145 / 952	provided simulations
Variables:	818795 / 93247	provided variables



Looking Back: Long-Term Archival for CMIP5 (2)

- Reason for Long-term archival (LTA) and the IPCC DDC (Data Distribution Centre) is to **provide stable data for long-term interdisciplinary use**:
  - Permanent and persistent access to stable data
  - of high-quality and
  - well-documented.
- LTA and IPCC DDC in CMIP5 were **no integral parts of the CMIP data infrastructure**.

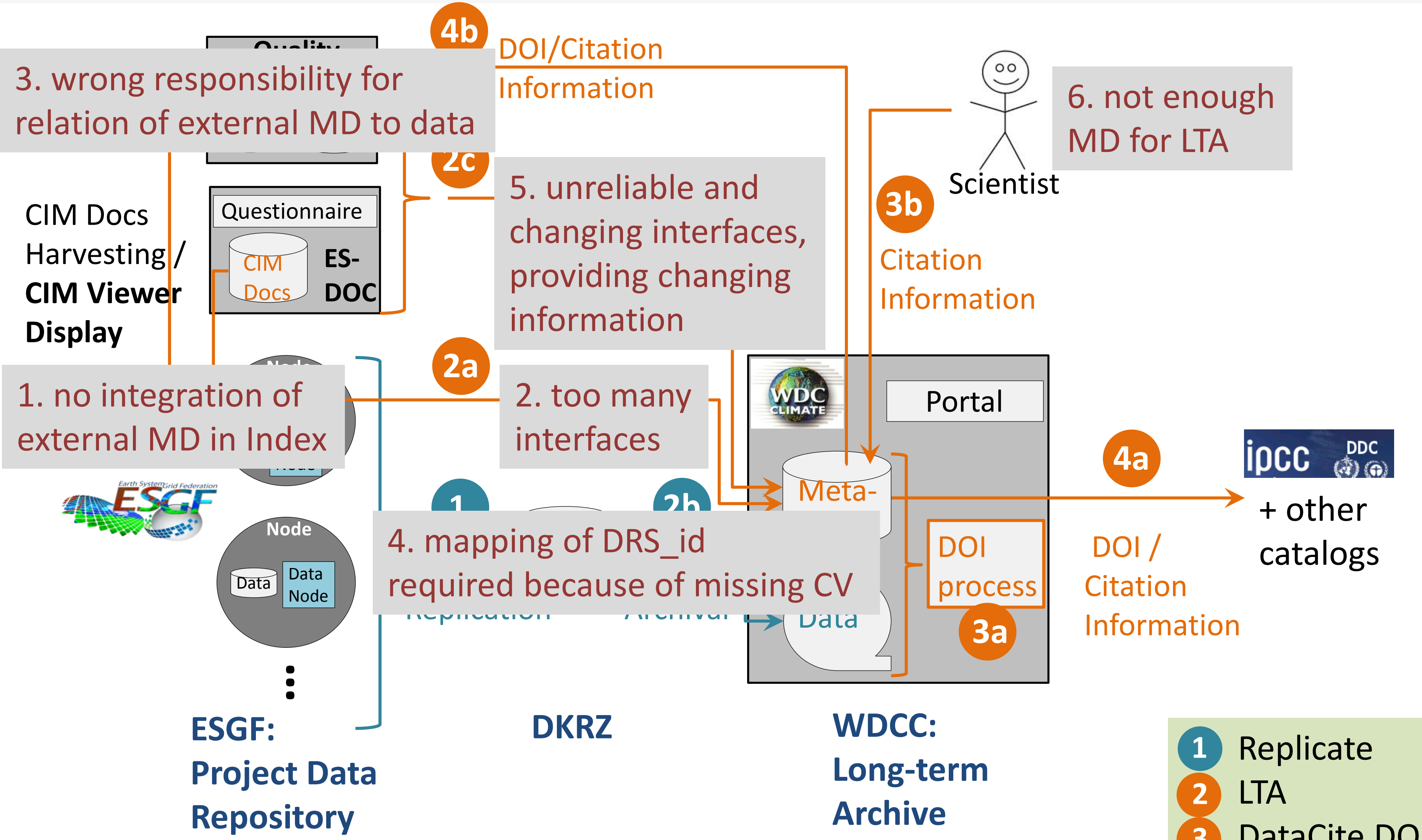
Looking Back: LTA Workflow in CMIP5



1	Replicate
2	LTA
3	DataCite DOI
4	Share DOI

Source: Stockhause, Martina. (2014). Long-term archiving workflow in CMIP5 - a first review. ISENES2 Workshop, Hamburg, 03.-05.06.2014. Zenodo. doi:10.5281/zenodo.29104

Looking Back: LTA Workflow in CMIP5 (3)



3. wrong responsibility for relation of external MD to data

6. not enough MD for LTA

5. unreliable and changing interfaces, providing changing information

1. no integration of external MD in Index

2. too many interfaces

4. mapping of DRS_id required because of missing CV

1 Replicate
2 LTA
3 DataCite DOI
4 Share DOI

Source: Stockhause, Martina. (2014). Long-term archiving workflow in CMIP5 - a first review. ISENES2 Workshop, Hamburg, 03.-05.06.2014. Zenodo. doi:10.5281/zenodo.29104

II. Long-Term Archival (LTA) in CMIP6

LTA Perspective of CMIP6

Expected values for CMIP6 (CMIP5 values) :

- Volume of CMIP6 data: 10-90 PBytes (2 PBytes)
- Volume of AR6 data: 2-3 PBytes (1.6 PBytes)
- Number of Data Nodes: 25 (17)
- Number of Metadata Repositories: 5 (2)

→AR6 will be a subset of the CMIP6 snapshot

→Integration of metadata from repositories need to be better organized

Looking Back: Recommendation from 2014 (1)

- **Project administration:** **WGCM Infrastructure Panel (WIP)**
  - (✓) *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
  - **CMIP Data Pool** 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.

Looking Back: Recommendation from 2014 (2)

- **CMOR2:** **furtherInfoURL (CIM/Citation) and PIDs**

(✓)

  - 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:**

Core Data Nodes

  - *Enforcement* of consistent use of identifiers and data versioning and other agreed *policies*

✓

  - 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)

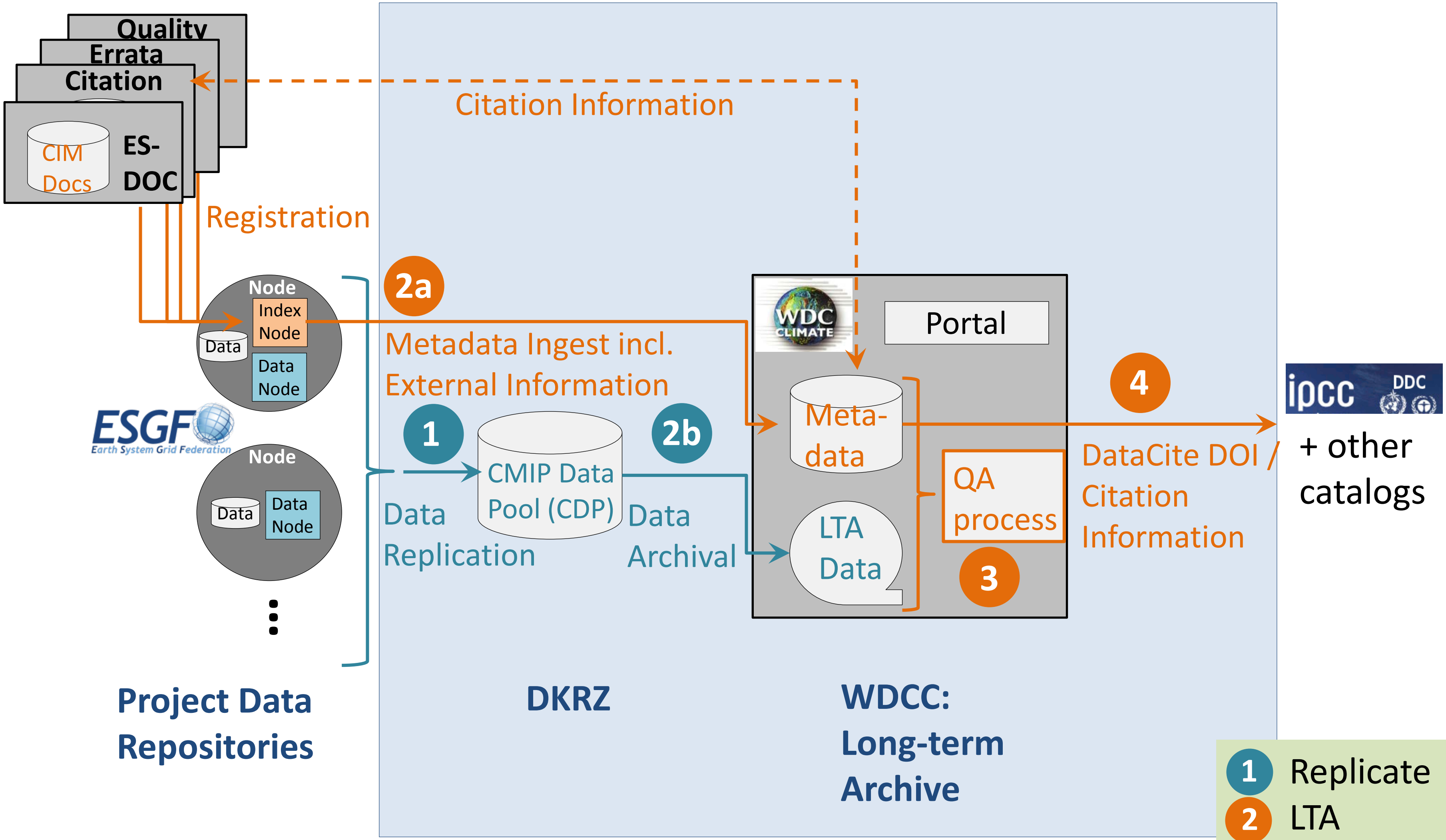
Ancillary Metadata

  - *Integration of additional metadata* into ESGF, e.g. searchable selected CIM/Quality/Citation/Errata Annotation/Provenance metadata

Looking Back: Recommendation from 2014 (3)

- **Citation:** **CMIP6 Citation Service**
  - ✓ collect data citation information with the data, ideally with PID assignment
  - integration in reference lists of scientific papers

LTA Workflow for CMIP6



1 Replicate
2 LTA
3 DataCite DOI
4 Share DOI

Based on: Stockhause, Martina (2014). Long-term archiving workflow in CMIP5 - a first review. ISENES2 Workshop, Hamburg, 03.-05.06.2014. Zenodo. doi:10.5281/zenodo.29104

Long-Term Archival Improvements for CMIP6 (1)

1. LTA has become a part of the CMIP data infrastructure: WGCM Infrastructure Panel (WIP) white paper available: http://doi.org/10.5281/zenodo.35178
2. CV on DRS components available: https://github.com/WCRP-CMIP/CMIP6_CVs
    - No mapping of DRS components required
3. Registration of ancillary metadata in ESGF:
    - LTA has only to deal with metadata format but is no longer responsible for its connection to the data
4. CMIP6 Citation Service collects citation and contact information during CMIP6 (http://cmip6cite.wdc-climate.de):
    - No need for data provider to fill in the gaps in the metadata

III. IPCC Data Distribution Centre for AR5

IPCC DDC at WDCC / DKRZ

World Data Center for Climate

Long-Term Archive for Climate Data

- 1992: Long-term archive for climate data
- 2003: regular member of the ICSU World Data System, 2011 renewed ICSU WDS membership/certification
- 2010: WDCC moved to Deutsches Klimarechenzentrum

IPCC DDC at WDCC / DKRZ

Reference Archive for Climate Model Output Data

- 1995: LTA for IPCC climate model data since SAR
- 2008: parts of FAR added to DDC
- 2013/14: LTA of IPCC AR5

IPCC DDC: Reference Data Archive

The IPCC DDC provides data on the long-term for an interdisciplinary user community in support of the IPCC Authors.

Long-term:
archival with second data copy in an established data center

Interdisciplinary Use:
add information to the data for a creator-independent usage

IPCC Author Support:
provide a reliable, up-to-date and easily-accessible CMIP data pool

Experiences with IPCC Author Support in AR5

- CMIP5 data infrastructure was under development during data distribution:
  - Missing version management
    → intransparent data changes
  - Complicated authentication/authorization solution
    → data access barrier
  - Script-based access under development and not matching user requirements

→ ETH Zurich set up and managed a data repository to support the work of the IPCC WG1 authors

→ IPCC DDC long-term archived two data collections for AR5

- No communication between IPCC WGs and IPCC DDC/TGICA

IV. IPCC Data Distribution Centre for AR6

IPCC AR6: Timeframe

Updated timeframe for IPCC AR6 from 16th September 2016 (http://ipcc.ch/activities/pdf/ar6_schedule.pdf):

- 05/2017: AR6 Scoping Meeting
- 09/2017: IPCC approval of AR6 outline
- **02/2018: Decision on selection of authors**
- 05/2019: WG I AR6 first-order draft expert review
- **03/2020: WG I AR6 second-order draft expert review**
- 04/2021: WG I AR6 IPCC acceptance/adoption/approval at IPCC-53

IPCC DDC AR6: Plans for AR6

The CMIP6 data infrastructure will be improved in version management but not in AA solution and script-based access. Therefore a CMIP6 Data Pool is still required in support IPCC authors:

→ IPCC DDC can offer to open its CMIP data pool for IPCC authors:

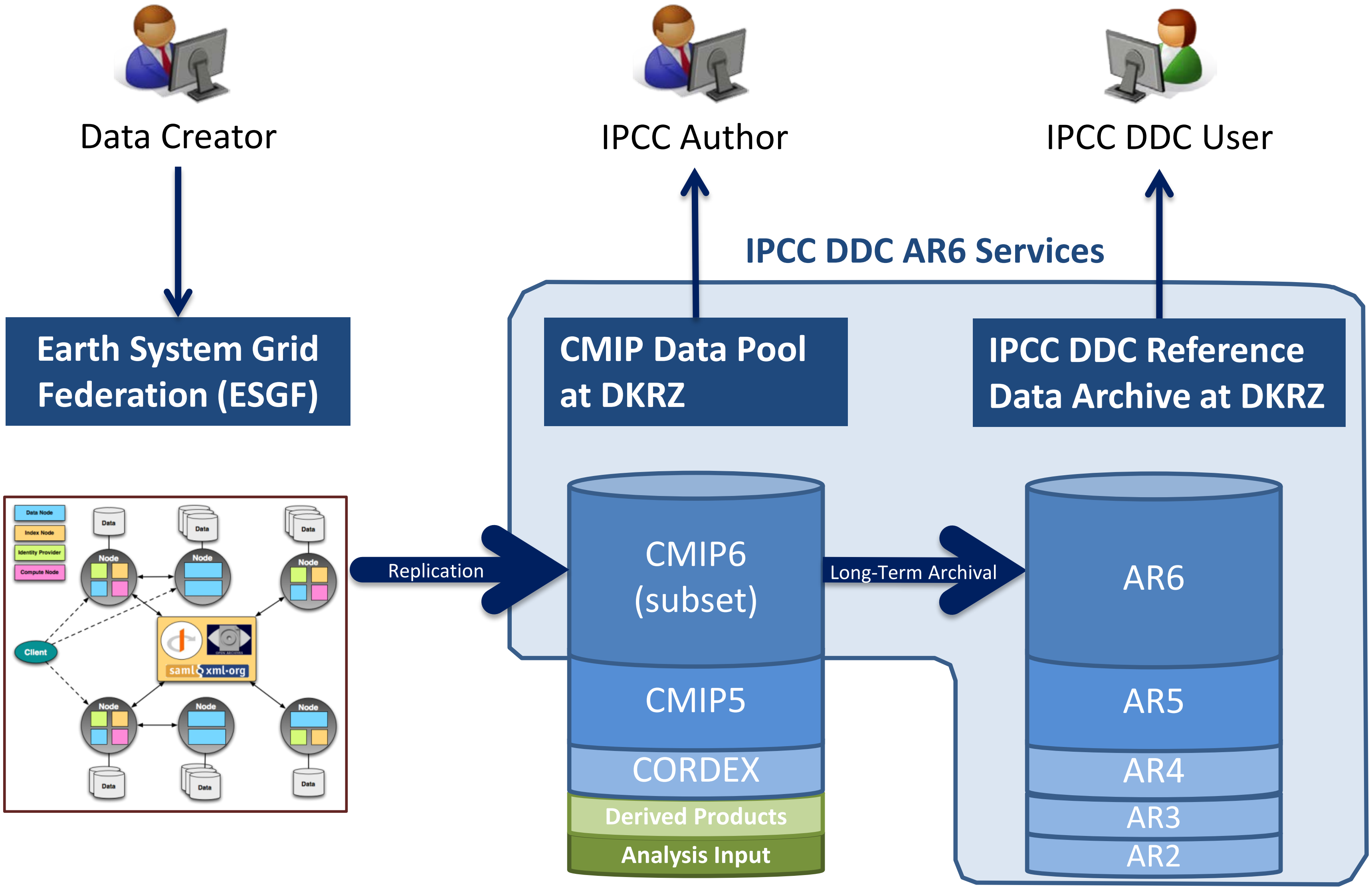
- Idea presented at “IPCC Expert Meeting on the future of TGICA” (Geneva, 01/2016 - IPCC-XLIII/Doc. 10, Corr.1)
- Information and Draft Concept available at https://redmine.dkrz.de/projects/dkrz_cdp/wiki
- Coordination between IPCC WG and IPCC DDC on Data Pool to be organized

→ IPCC DDC AR6 Reference Data Archive will consist of a single data collection based on the Data Pool

A better integration of IPCC DDC / TGICA into IPCC processes is needed:

- Under discussion

IPCC DDC Services for AR6



Data Creator

IPCC Author

IPCC DDC User

IPCC DDC AR6 Services

Earth System Grid Federation (ESGF)

CMIP Data Pool at DKRZ

IPCC DDC Reference Data Archive at DKRZ

Replication

Long-Term Archival

CMIP6 (subset)

CMIP5

CORDEX

Derived Products

Analysis Input

AR6

AR5

AR4

AR3

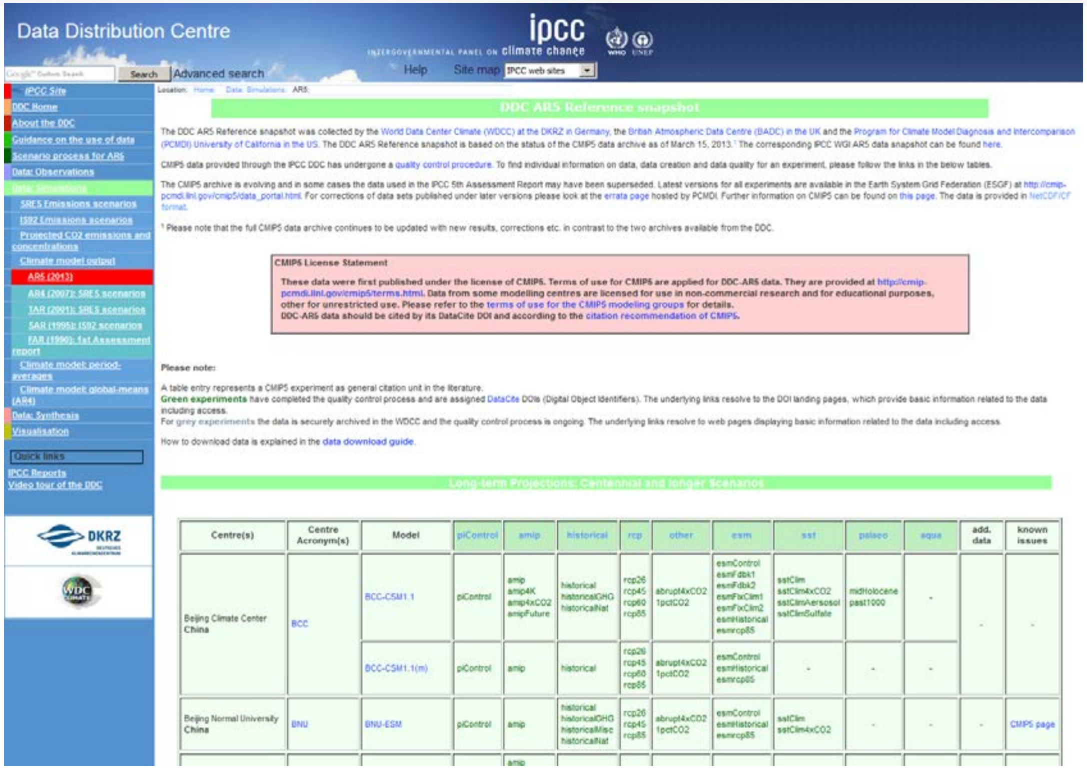
AR2

Long-Term Archival Improvements for CMIP6 (2)

1. LTA has become a part of the CMIP data infrastructure
2. CV on DRS components available
3. Registration of ancillary metadata in ESGF
4. CMIP6 Citation Service collects citation and contact information during CMIP6
5. DKRZ’s CMIP data pool is opened to IPCC authors:
   - AR6 data archival of this well-defined CMIP6 data subset used by the IPCC WGs

Summary

- Experiences from the CMIP5 workflow used to improve the LTA workflow for CMIP6
- Ad-hoc CMIP5 LTA workflow is transformed into a CMIP6 LTA concept
- LTA and IPCC DDC become part of the CMIP6 infrastructure gives the LTA a voice within the CMIP6 data infrastructure development
- IPCC author support will be strengthened



Data Distribution Centre

INTERGOVERNMENTAL PANEL ON climate change

Search Advanced search Help Site map IPCC web sites

- IPCC Site
- DDC Home
- About the DDC
- Guidance on the use of data
- Scenario process for AR5
- Data: Observations
- Data: Simulations
  - SRES Emissions scenarios
  - IS92 Emissions scenarios
  - Projected CO2 emissions and concentrations
  - Climate model output
    - AR5 (2013)
    - AR4 (2007): SRES scenarios
    - TAR (2001): SRES scenarios
    - SAR (1995): IS92 scenarios
    - FAR (1990): 1st Assessment report
  - Climate model: period-averages
  - Climate model: global-means (AR4)
- Data: Synthesis
- Visualisation

Quick links

IPCC Reports
Video tour of the DDC

Location: Home Data: Simulations AR5

DDC AR5 Reference snapshot

The DDC AR5 Reference snapshot was collected by the World Data Center Climate (WDCC) at the 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. The DDC AR5 Reference snapshot is based on the status of the CMIP5 data archive as of March 15, 2013.[1] The corresponding IPCC WGI AR5 data snapshot can be found here.

CMIP5 data provided through the IPCC DDC has undergone a quality control procedure. To find individual information on data, data creation and data quality for an experiment, please follow the links in the below tables.

The CMIP5 archive is evolving and in some cases the data used in the IPCC 5th Assessment Report may have been superseded. Latest versions for all experiments are available in the Earth System Grid Federation (ESGF) at http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html. For corrections of data sets published under later versions please look at the errata page hosted by PCMDI. Further information on CMIP5 can be found on this page. The data is provided in NetCDF/CF format.

[1] Please note that the full CMIP5 data archive continues to be updated with new results, corrections etc. in contrast to the two archives available from the DDC.

CMIP5 License Statement

These data were first published under the license of CMIP5. Terms of use for CMIP5 are applied for DDC-AR5 data. They are provided at http://cmip-pcmdi.llnl.gov/cmip5/terms.html. Data from some modelling centres are licensed for use in non-commercial research and for educational purposes, other for unrestricted use. Please refer to the terms of use for the CMIP5 modeling groups for details.
DDC-AR5 data should be cited by its DataCite DOI and according to the citation recommendation of CMIP5.

Please note:

A table entry represents a CMIP5 experiment as general citation unit in the literature.
Green experiments have completed the quality control process and are assigned DataCite DOIs (Digital Object Identifiers). The underlying links resolve to the DOI landing pages, which provide basic information related to the data including access.
For grey experiments the data is securely archived in the WDCC and the quality control process is ongoing. The underlying links resolve to web pages displaying basic information related to the data including access.

How to download data is explained in the data download guide.

Long-term Projections: Centennial and longer Scenarios

Centre(s)	Centre Acronym(s)	Model	piControl	amip	historical	rcp	other	esm	sst	palaeo	aqua	add. data	known issues
Beijing Climate Center China	BCC	BCC-CSM1.1	piControl	amip amip4K amip4xCO2 amipFuture	historical historicalGHG historicalNat	rcp26 rcp45 rcp60 rcp85	abrupt4xCO2 1pctCO2	esmControl esmFdbk1 esmFdbk2 esmFixClim1 esmFixClim2 esmHistorical esmrcp85	sstClim sstClim4xCO2 sstClimAerosol sstClimSulfate	midHolocene past1000	-	-	-
		BCC-CSM1.1(m)	piControl	amip	historical	rcp26 rcp45 rcp60 rcp85	abrupt4xCO2 1pctCO2	esmControl esmHistorical esmrcp85	-	-	-		
Beijing Normal University China	BNU	BNU-ESM	piControl	amip	historical historicalGHG historicalMisc historicalNat	rcp26 rcp45 rcp85	abrupt4xCO2 1pctCO2	esmControl esmHistorical esmrcp85	sstClim sstClim4xCO2	-	-	-	CMIP5 page
				amip									

Data Citation CMIP6: http://cmip6cite.wdc-climate.de

IPCC DDC: http://ipcc-data.org

DDC at DKRZ: http://ipcc.wdc-climate.de

M. Stockhause, F. Toussaint, M. Lautenschlager (2015): CMIP6 Data Citation and LTA. WIP white paper. Zenodo. doi:10.5281/zenodo.35178.