DKRZ

Data center requirements and services

Stephan Kindermann, Michael Lautenschlager, Katharina Berger,
Tobias Weigel, Hans Dieter Hollweg
Deutsches Klimarechenzentrum (DKRZ)

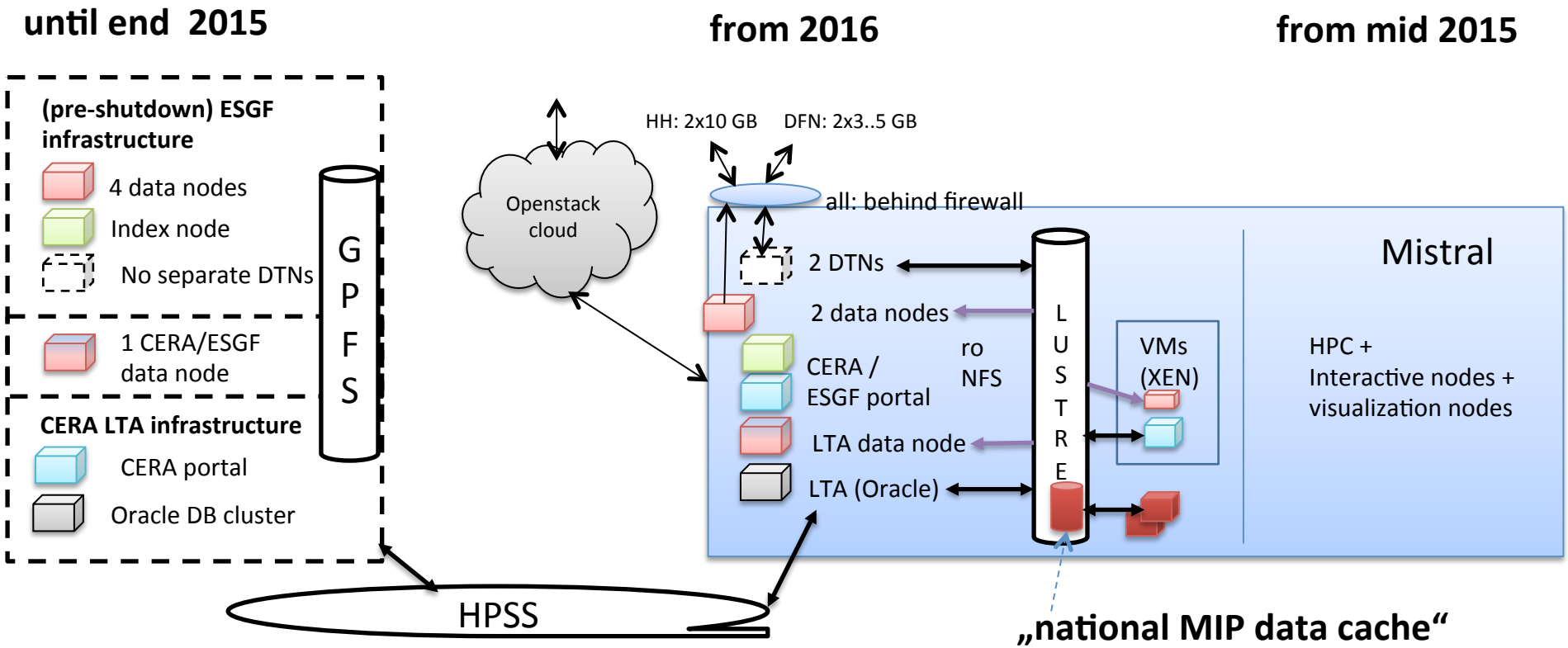
Overview

- Update: the new data infrastructure hosting environment at DKRZ
- ESGF: DKRZ data life cycle services
  - LTA / WDCC – ESGF integration
  - Quality assurance
  - Data near processing
  - Towards PID based services
- CMIP6 at DKRZ

DKRZ data center update

Migration to new integrated HPC / data system

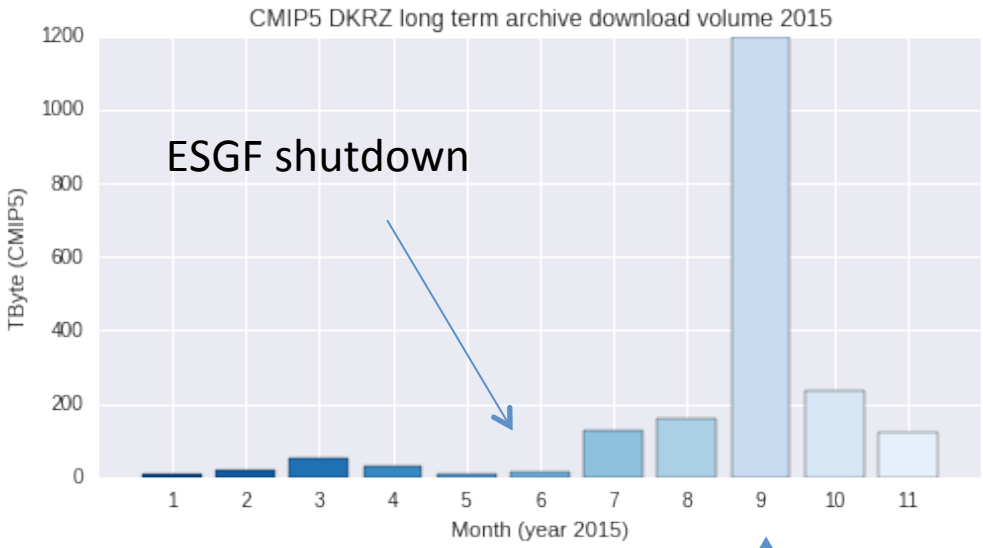
- separate DTNs (starting 2016)
- establishment of a „national MIP data analysis cache“
- data cloud to support data ingest process

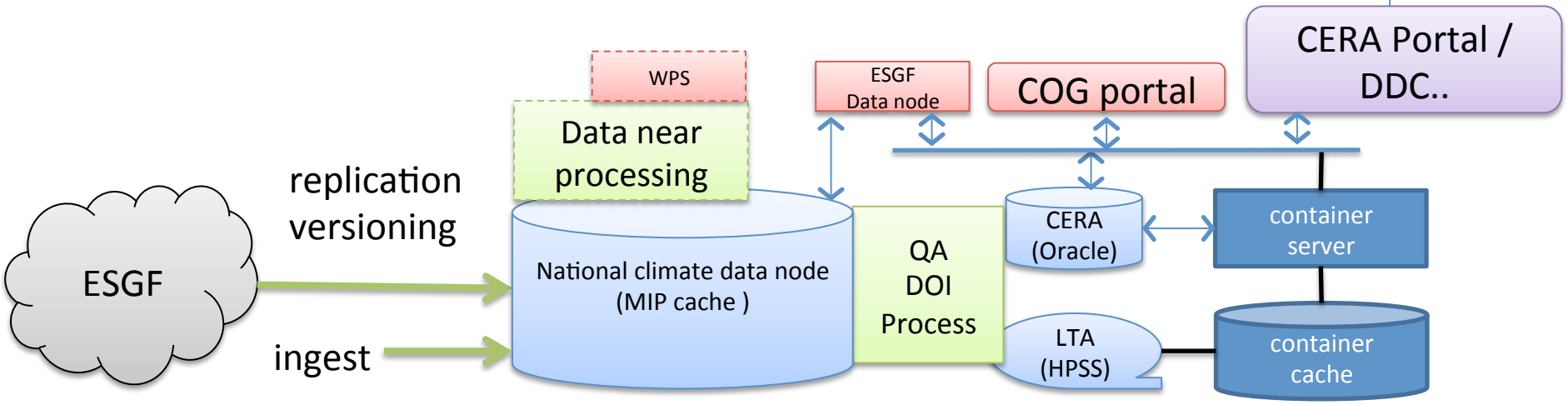


DKRZ long term archival and data citation

Mayor use case

- Replication
- **Support data evaluation**
- **Quality Assurance**
- Long Term Archival
- DOI assignment
- **Exposure as ESGF data node**





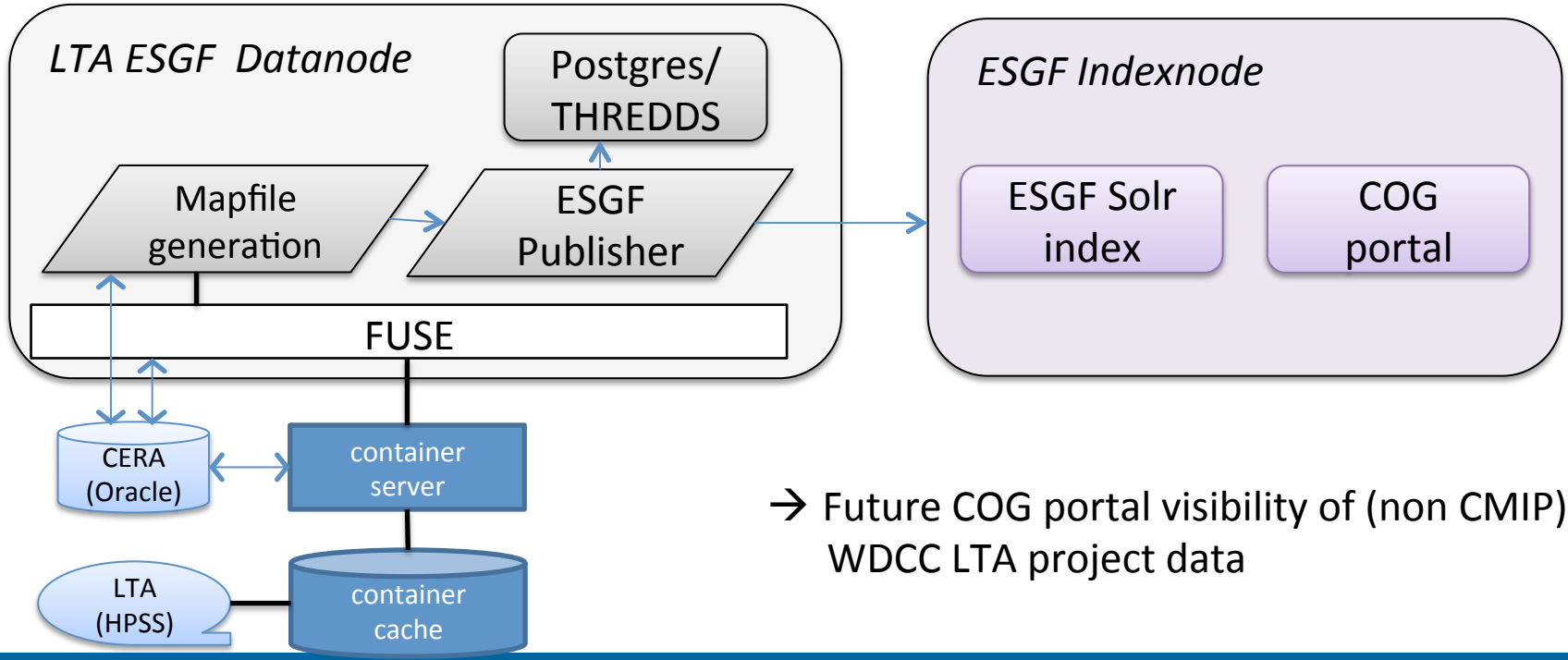
WDCC / CERA / HPSS ←→ ESGF integration

Operational for CMIP5

- CERA metadata (Oracle) → ESGF index
- Thredds server with ESGF security filter + HPSS data container server → ESGF data node

Improved system for CMIP6:

- FUSE based mounting of DKRZ HPSS/cache legacy system
- Extraction of CERA metadata for ESGF mapfile
- „standard“„standard“ ESGF publication in an „offline mode“

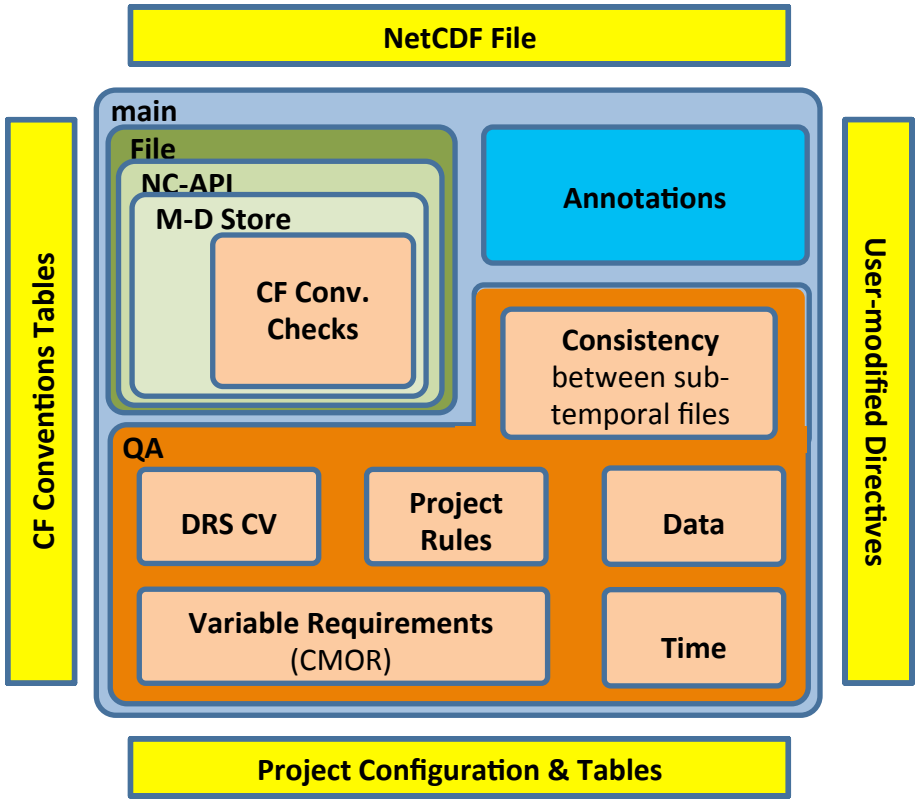


→ Future COG portal visibility of (non CMIP) WDCC LTA project data

(CMIP data) Quality Assurance Software

Completely re-structured and modularized:

- Flexible configuration
- Used heavily for CORDEX – will support CMIP6
- Separate cf-checker module



CF Conventions Check

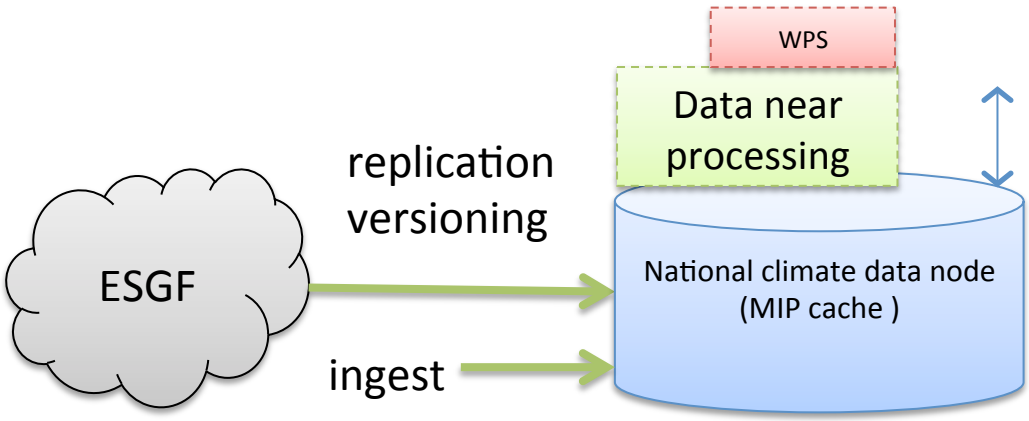
- Versions: 1.4 - 1.6
- 8-9 Chapters of rules
- table based config (area-type, cf-standard-name, stand-region-name, ..)

- **Source code:** https://github.com/h-dh/QA-DKRZ
- **Pre-packaged versions:** conda based, docker based
- **Documentation:** http://qa-dkrz.readthedocs.org/en/latest/qa-user-manual.html

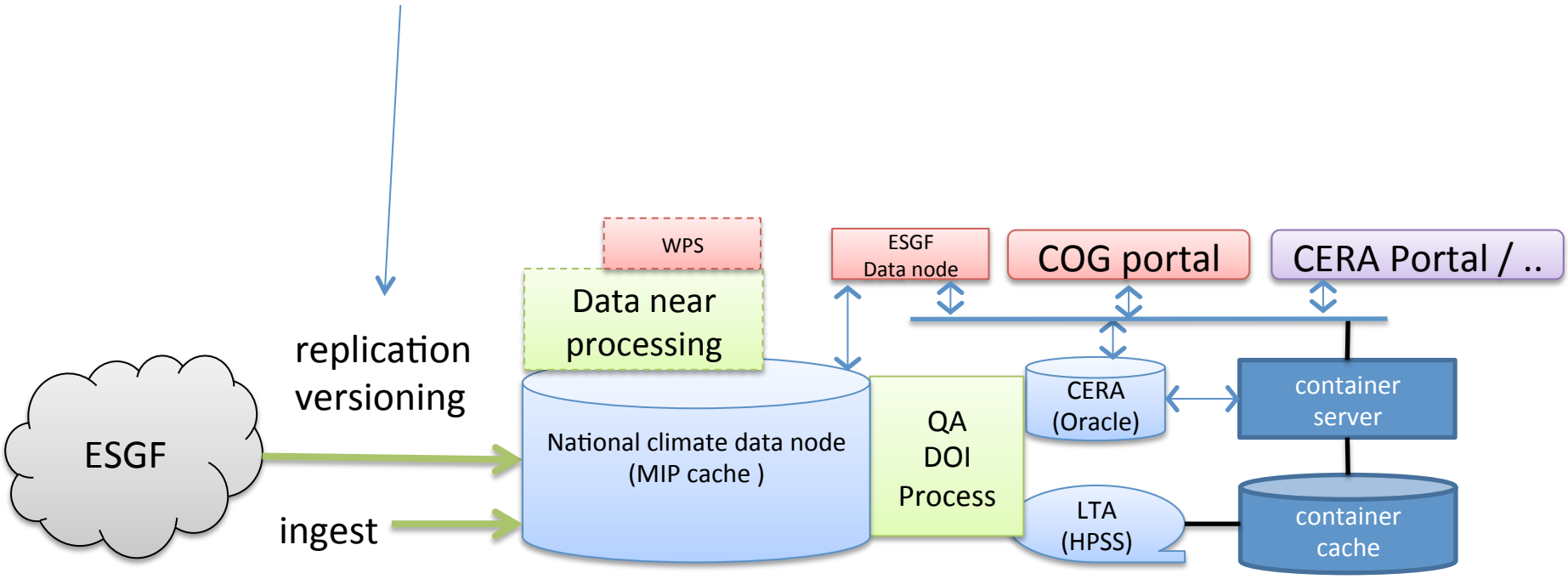
National MIP data analysis cache / node

„Ad hoc“ approach → transparent solution:

- Data needed → help desk → data manager
- RO mounted on HPC data analysis nodes
- Support for data analysis VM deployment
- Support for tool dependency management (install recipes, conda, docker)
- WPS framework to support web service deployments
  - Birdhouse (https://github.com/bird-house)
  - conda/docker support
  - Support for home institution (test-) deployments



Stable file/collection management !?



Towards PID based services

Motivation: Stable ESGF data space based on PID infrastructure

Collaborations:

- ePIC: DKRZ partner → prefix registration
- EUDAT: DKRZ leads PID task → API
- RDA: DKRZ co-chairs PIT and collections WGs
- Envri+: PIDs in environmental sciences

Next ESGF steps:

- Test-Environment (PID system + publisher)
- Scalable, stable PID assigment:
  - CMOR integration, CDNOT involvement
  - PID API / ESGF publisher integration
  - High available message queuing system integration

Summary

Long term archival use case

→ ESGF integration

→ Quality Assurance

→ PID assigment early in data life cycle

→ early citation and DOI assignment

→ future PID based data management services

→ future PID based end user services

→ future PID based provenance support

Thank You

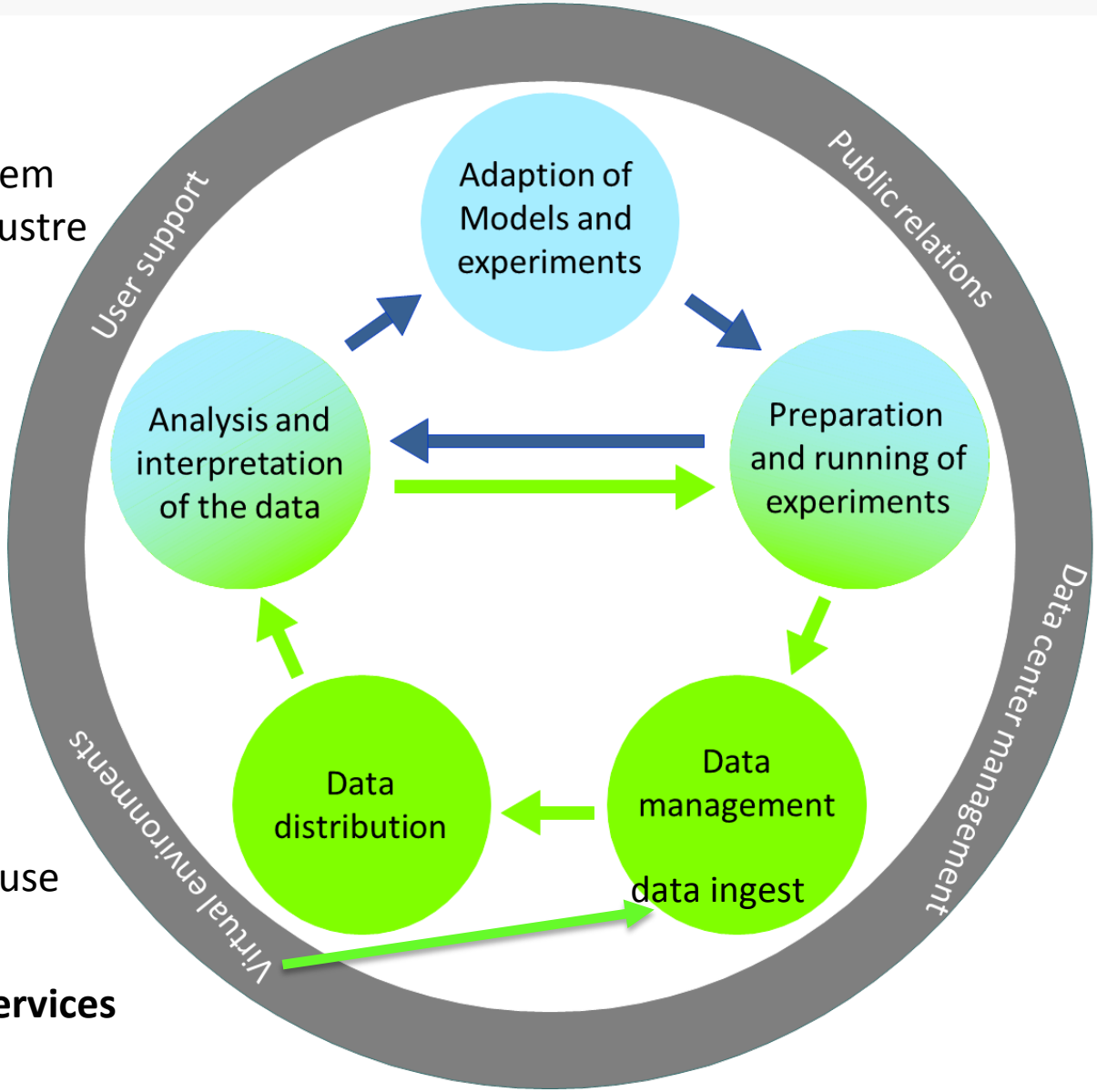
DKRZ services

New developments

- New integrated HPC/Data System installed in 2015, ~ 50 PByte Lustre
- Storage cloud (openstack)
- Community data analysis cache and platform

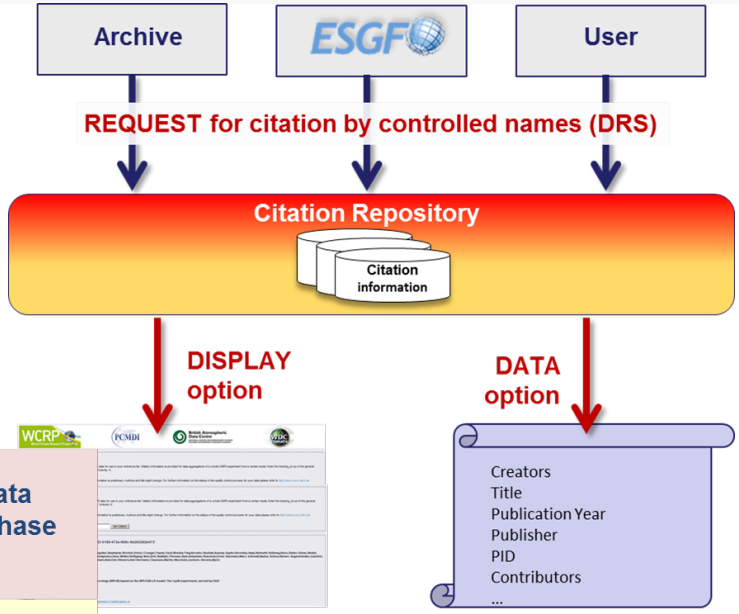
ESGF:

- WDCC/HPSS/ESGF data node
- WPS compute platform birdhouse
- **Towards PID / early citation services**



(Early) Data Citation (DM + ESGF)

- Impact on CMIP6 data management (DM) and ESGF governance (ESGF)
- Request from modelling groups for a data citation reference just after ESGF data publication
- CMIP6 data publication workflow:



Paper Publication	Data Publication	Data Workflow	
	Data partly available	ESGF publication of initial (incomplete) data	**Initial Data Sharing Phase**
Paper Preparation	Data Analyses		
Draft Paper Publication	Early Data Publication	ESGF publication of Revised Datasets as new versions and withdrawal of old versions	**Data Review Phase**
Paper Peer-Review	Data Community Review		
Final Paper Publication (Early Data Citations)			
Final Paper Publication (DataCite DOI Data Citations)	LTA Data Publication	Data is finalized, long-term archived, and DataCite DOI Published	**Stable Data Phase**

Time

CMIP6 citation granularities are collection levels:

- **Simulation**
- **Model**