Archivierung am DKRZ

Auf dem Weg zur Datenpublikation

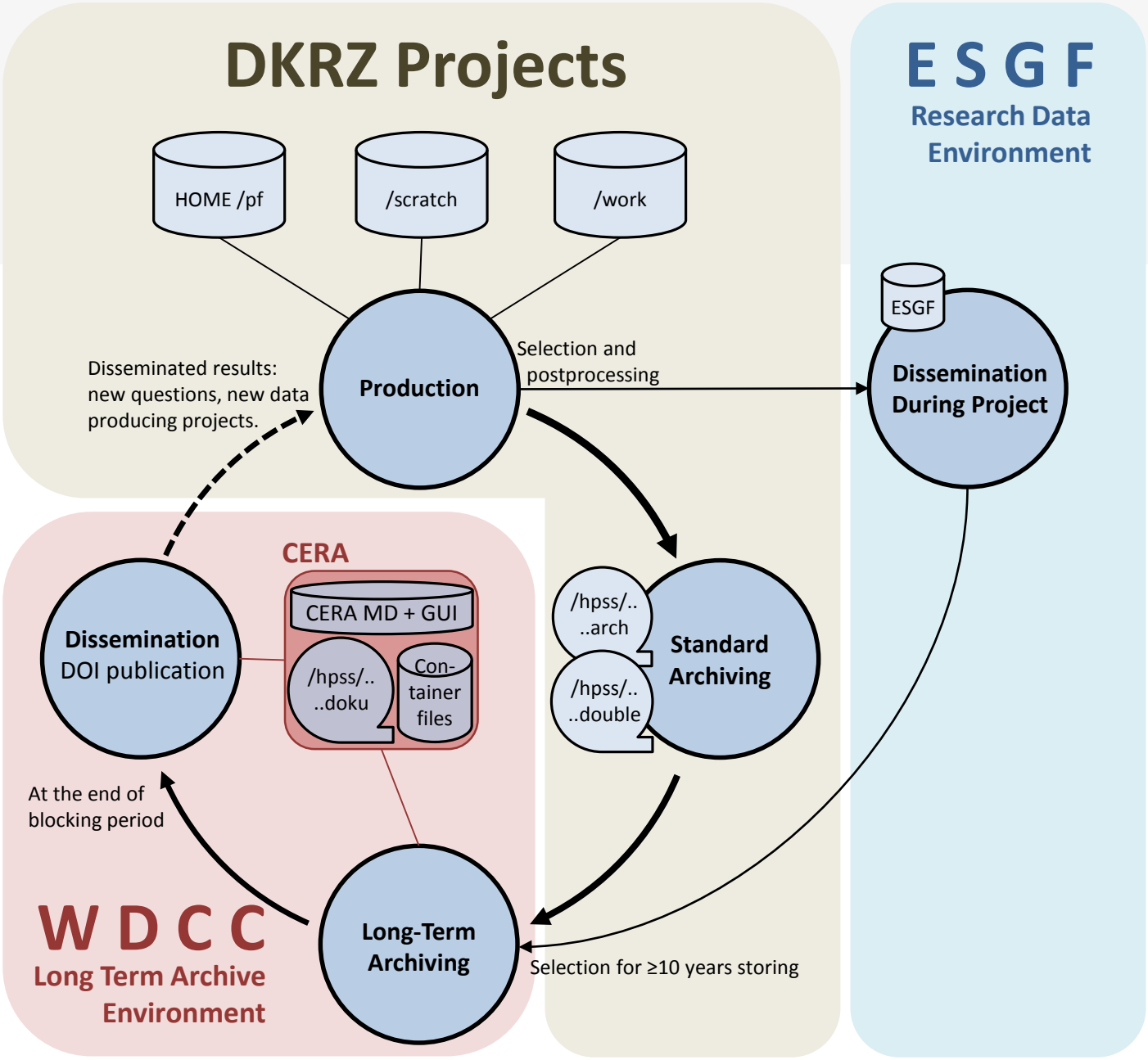
Workshop DKRZ-Datendienste
7. Mai 2014

Frank Toussaint
Deutsches Klimarechenzentrum (DKRZ)

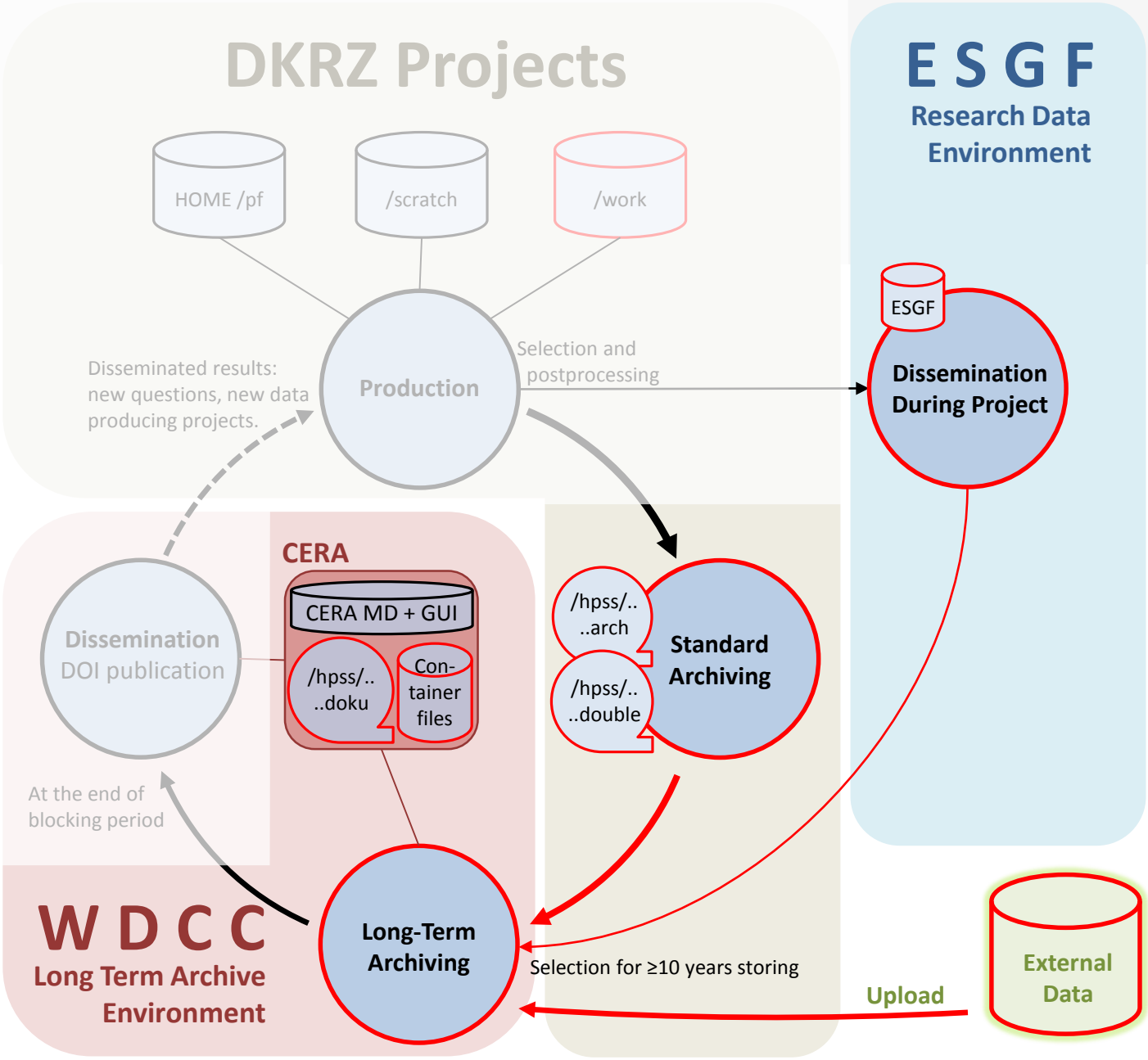
Contents

- Workflow to the Archive
- Data Management Plans
  - What to Describe
  - Example CORDEX
  - Cost Estimation
- ...and the Metadata?

Workflow to the Archive



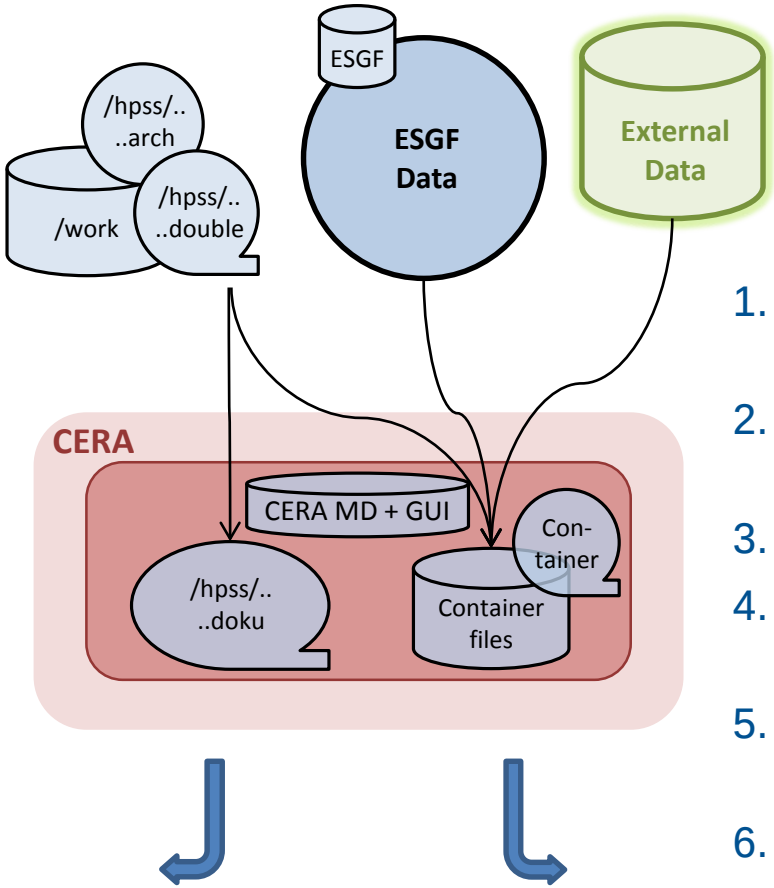
Workflow to the Archive



Workflow to the Archive

Minimum Archiving:
/hpss/doku

1. Data of projects at DKRZ
2. Access: DKRZ-User
3. Not designed for reuse by third parties
4. Just basic MD, maybe use of READMEs
5. No Digital Object Identifier (DOI)
6. Storage time = project + **10** years



Archiving for Reuse:
CERA Container Files

1. Also external data & ESGF data
2. Access: global, interdisciplinary
3. Reuse of data is possible
4. CERA-Metadata in the Database
5. With further qualification: DOI possible
6. Storage time >= project + **10+** years

Archiving for Reuse

Workflow to the Archive

1. Project manager contacts data@dkrz.de :
   - User name, institute's data (name, address, e-mail, phone)
   - Project name & contact (name, address, e-mail, phone)
   - List of files/datasets including single and total sizes
   - Further project information

2. Agreement on optimal data structures: proposals/advice by DKRZ, additional processing may be necessary (see below).

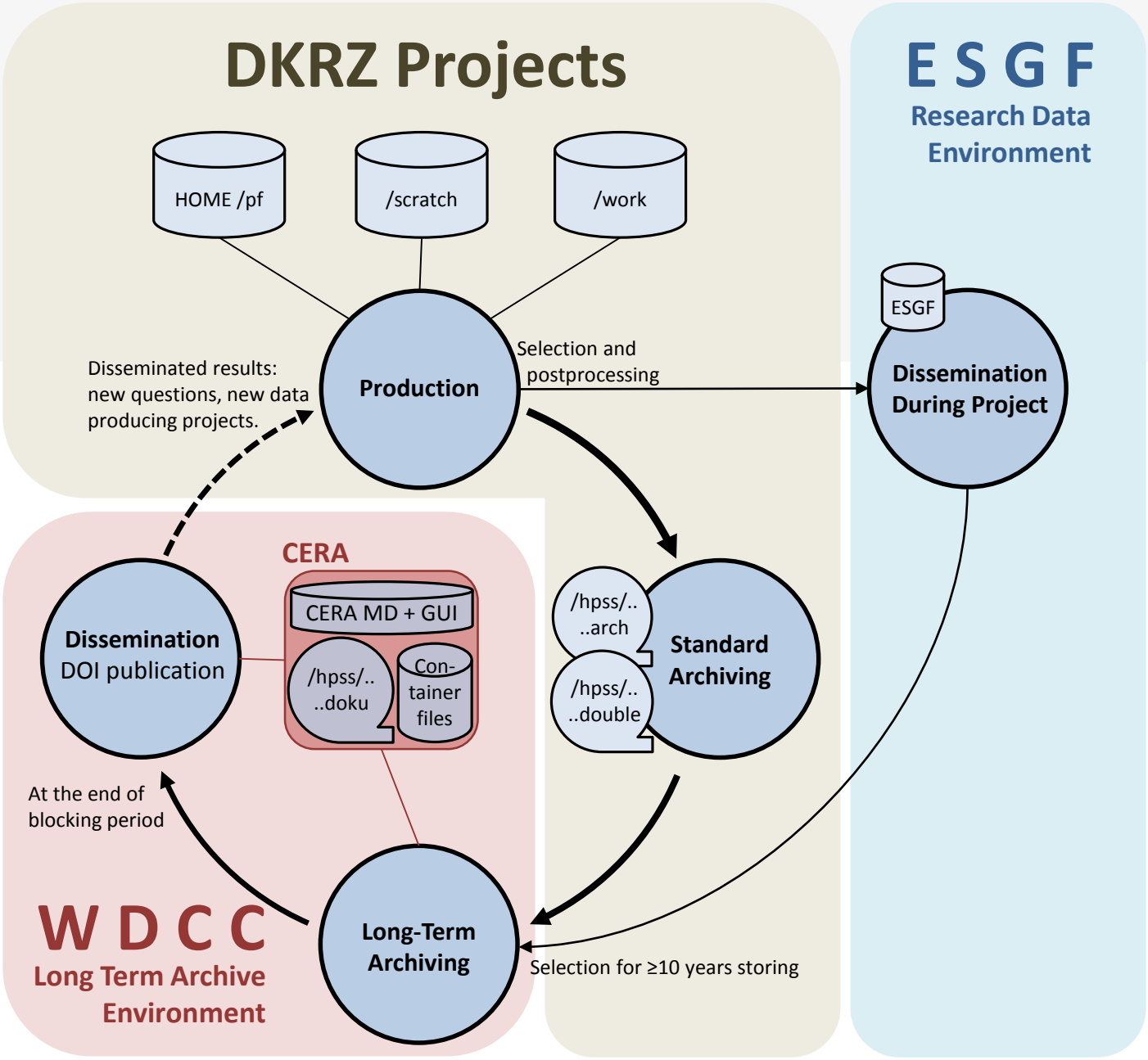
3. Insertion of metadata by the user (Login via CERA-Account) at:
   http://cera-www.dkrz.de/LTA_metadata
   Metadata need to be confirmed by DKRZ.

4. Additional data processing if needed.

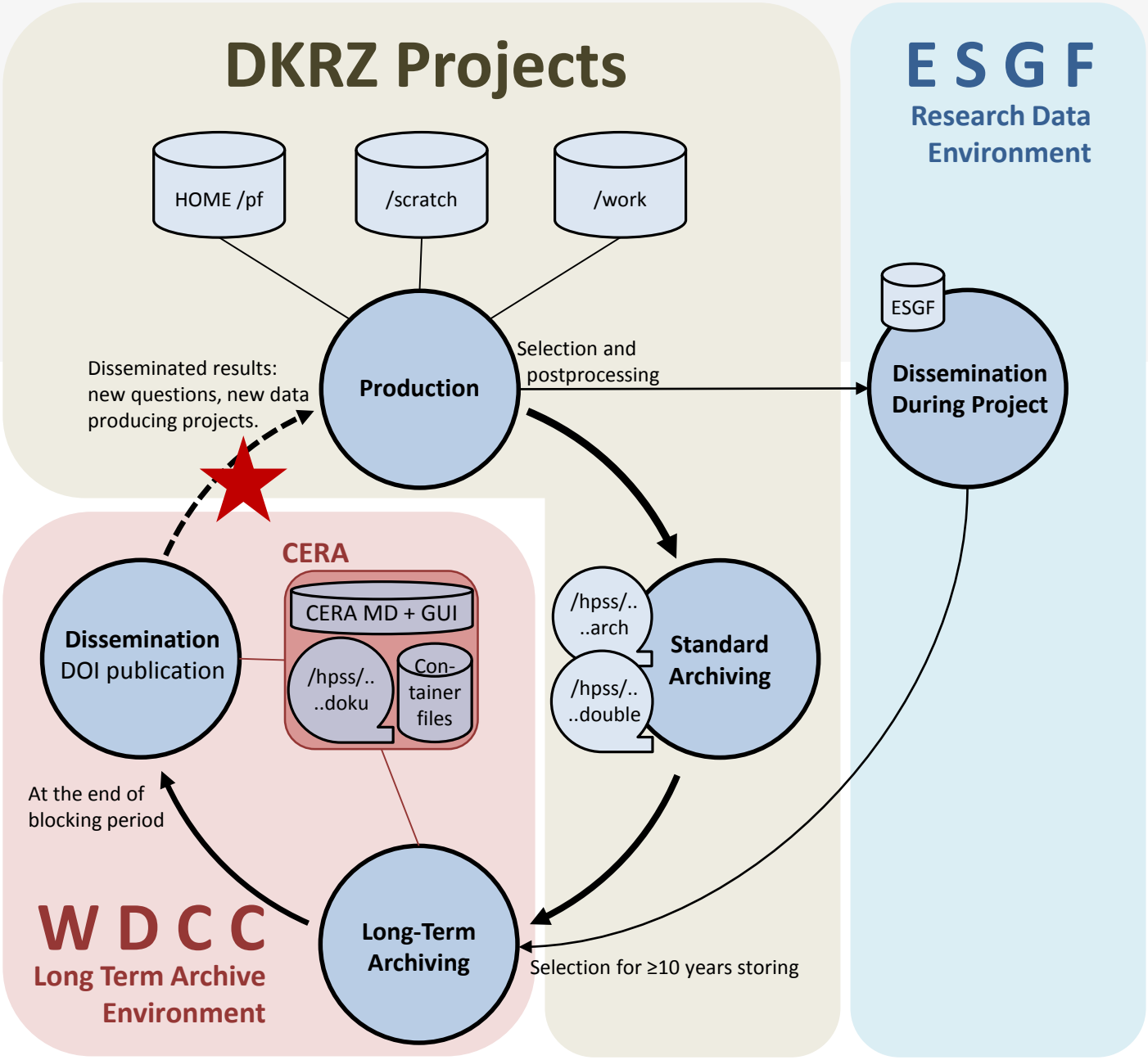
5. Upload of metadata
   Upload of data

6. Final checks of metadata and data by DKRZ (technical checks) and user (content checks) & publication.

Data Management Plans



Data Management Plans



Data Management Plans

...at application for funding!

Necessary Content:

- Data structures and workflows
- Data volumes, formates
- Data access: from when, by whom, until when, by which means?
- Publication? Responsable author?
- Further services needed

Data Management Plans

Example CORDEX:

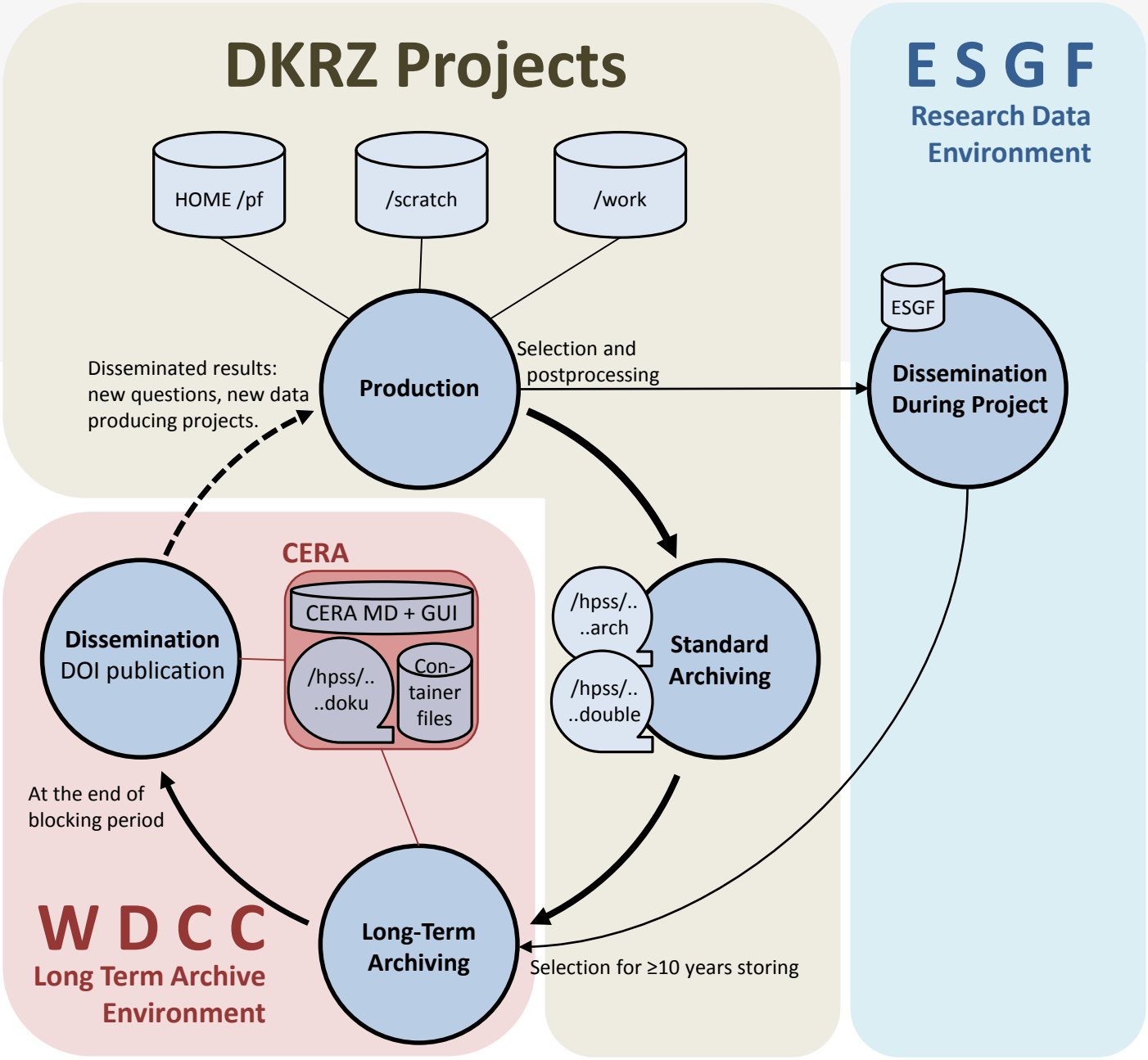
- Introduction – Overview – data storage and data services
- Archive Contents – Data and file format – Metadata in file headers – Variables – Submitting data – Data Volume
- Data nodes – Data node management – Publication units – User support – ESGF index nodes – Long term archiving (LTA)
- Long Term Archive at DKRZ – LTA at other CORDEX data centers
- DOI – local archives
- Data Quality Control (QC) – Levels of QC, Execution of QC
- Data access and sharing – Terms of use
- Annotation Phase – General Annotation Process
- Observational data – Backup policies
- Open issues – Action items – Responsibilities
- Glossary – Links and References

Cost Estimation

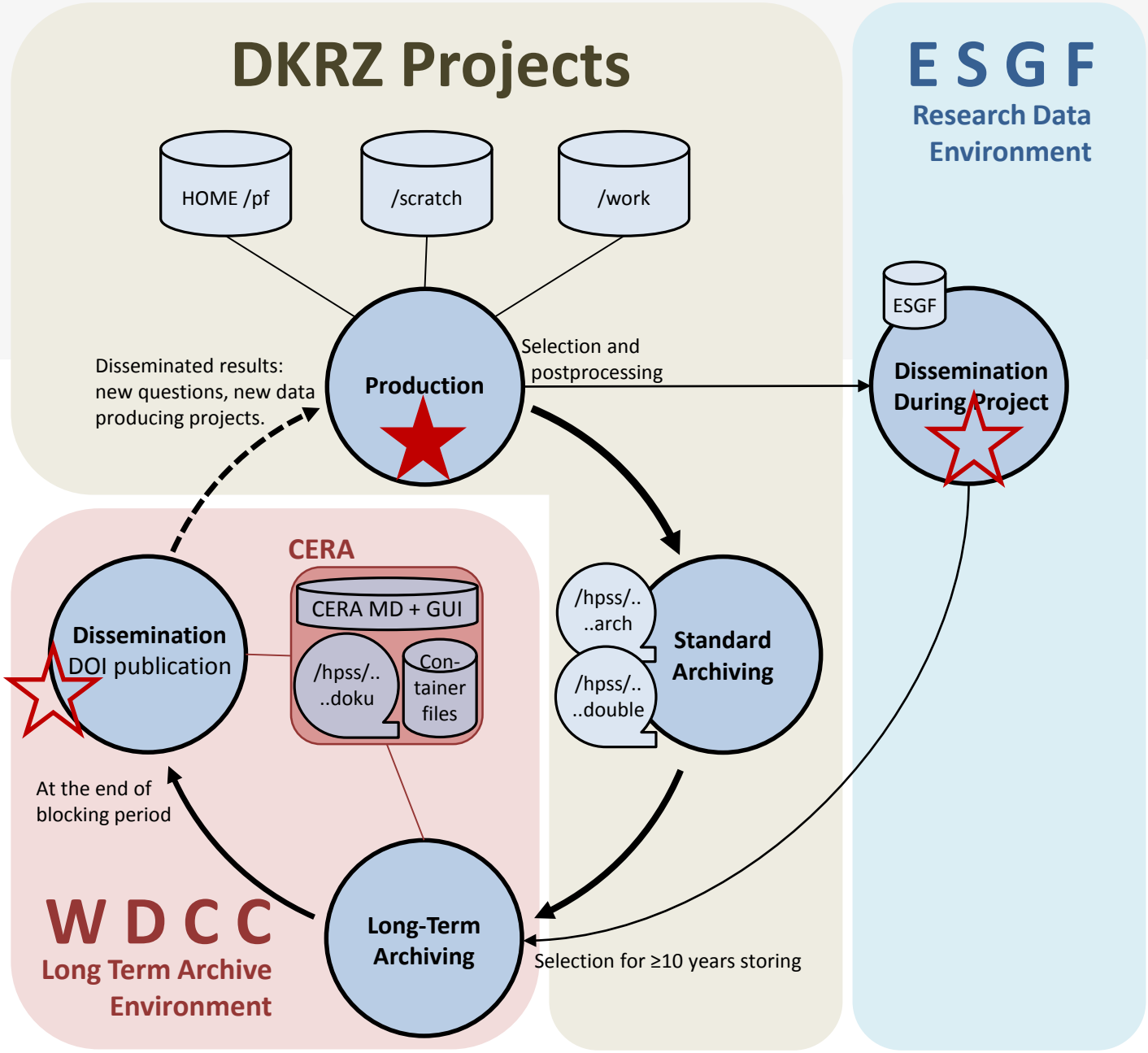
Cost prices:

- MD generation & storage: xml?, manually?
- Data storage: processing necessary?
- Costs of processing & ingest
- Costs of curation, media & operation
  - operation costs:
    - energy, staff, amortisation, network

...and the Metadata?



...and the Metadata?



...and the Metadata (MD)?

Why Metadata?

- For search in tens of PBytes – search MD
  - Different pre-conditions of different users
- For interpretation of data – use MD
- For reuse – rights, contact MD
- For DOI: Citability – citation MD

...and the Metadata?

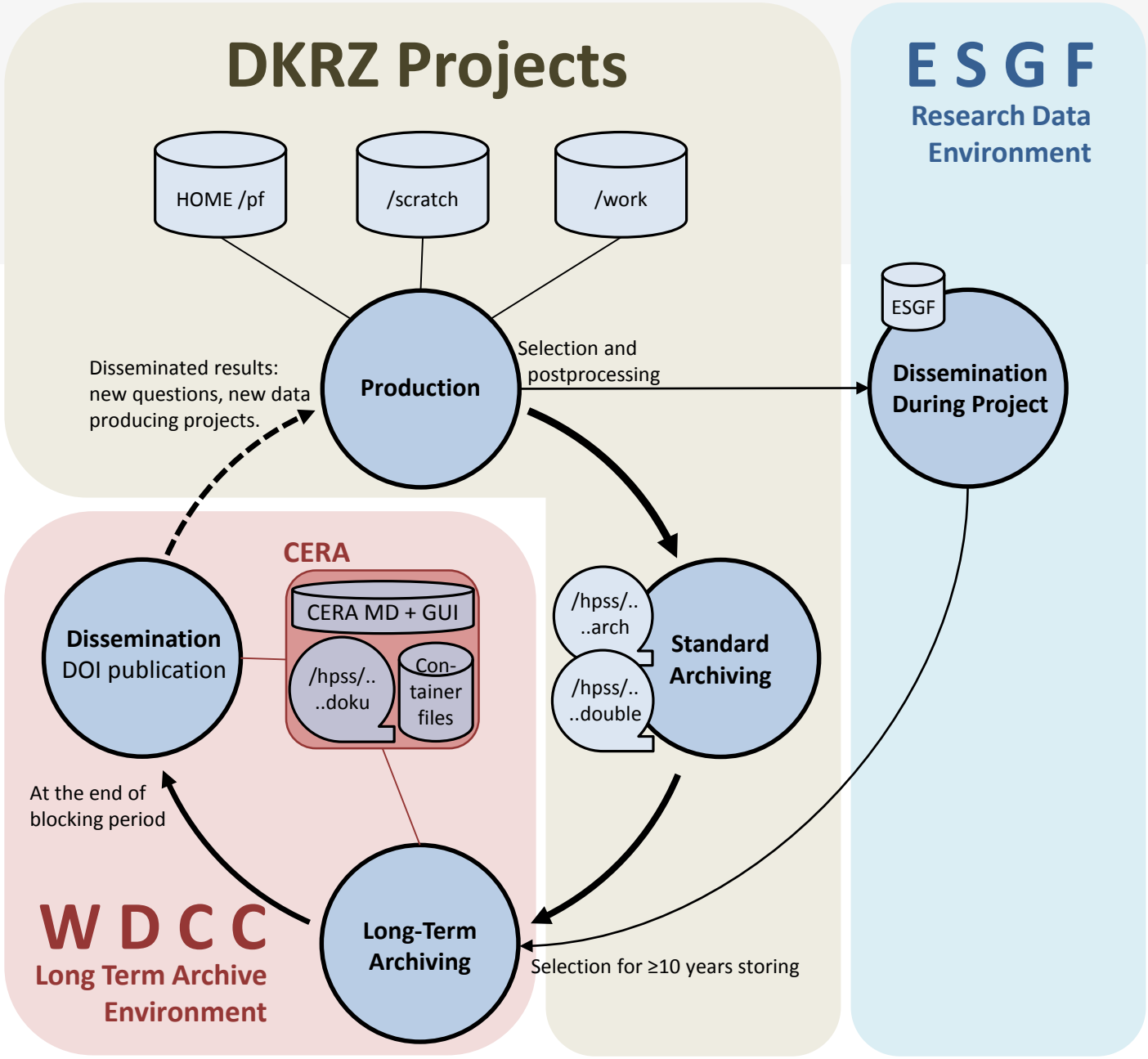
Where do Metadata Come from?

- Parameter lists and xml files from the projects
- Metadata collected from NetCDF file headers as in ESGF
- Coming (!) soon (?): metadata generation in the workflow
- Manuelle Eingabe von Metadaten (DOI-Publikation u.a.)

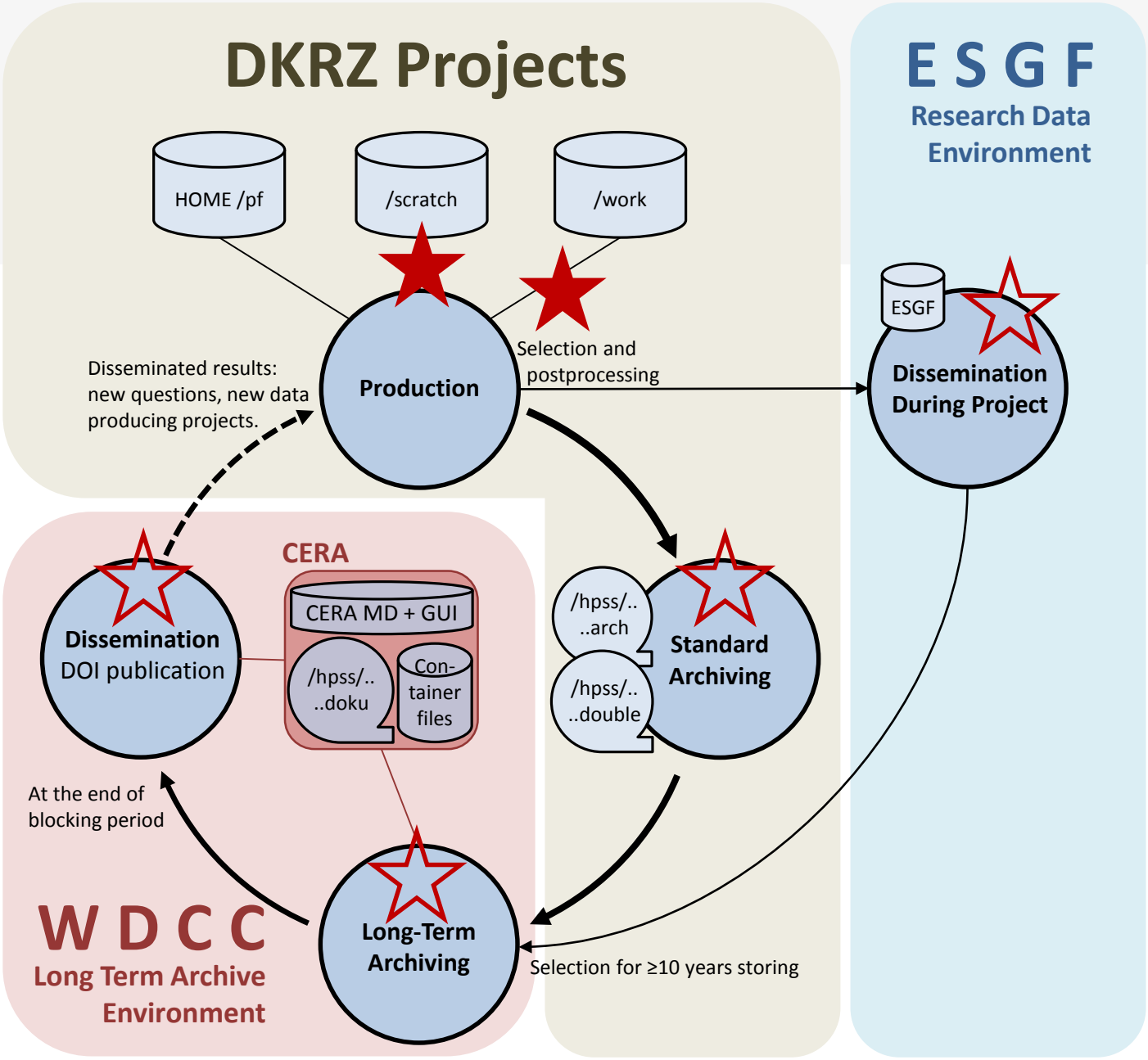
...and the Metadata?

Which Requirements and Standards for metadata?

- Minimum MD: name, project, contacts (PI, MD), variables, if applicable: coverage in space and time
- Normal: citation elements, further contacts, format, access rights, coordinate systems, ...
- DOI: results of quality checks, ...


DKRZ Projects
HOME /pf
/scratch
/work
Production
Disseminated results:
new questions, new data
producing projects.
Selection and
postprocessing
ESGF
Dissemination
During Project
CERA
Dissemination
DOI publication
CERA MD + GUI
/hpss/..
..doku
Con-
tainer
files
/hpss/..
..arch
/hpss/..
..double
Standard
Archiving
At the end of
blocking period
W D C C
Long Term Archive
Environment
Long-Term
Archiving
Selection for ≥10 years storing
E S G F
Research Data
Environment
P I D

PID


DKRZ Projects
HOME /pf
/scratch
/work
Production
Disseminated results:
new questions, new data
producing projects.
Selection and
postprocessing
ESGF
Research Data
Environment
Dissemination
During Project
CERA
Dissemination
DOI publication
CERA MD + GUI
/hpss/..
..doku
Con-
tainer
files
/hpss/..
..arch
/hpss/..
..double
Standard
Archiving
At the end of
blocking period
W D C C
Long Term Archive
Environment
Long-Term
Archiving
Selection for ≥10 years storing


VIELEN DANK !

...

...

...