some questions ….

Rise your hand if you agree …

Did you once find data on your or the institute's computer and no one knew any details about this data anymore because forgotten or the person who generated the data or put it there is no longer at the institute?

Rise your hand if you agree …

You once needed data from someone else, e.g. to compare your results, but perhaps also a script or part of model code. You may even have known the person personally .. BUT you had to wait forever for the data, were put off because of lack of time and got the data late or maybe not at all?

Rise your hand if you agree …

Or perhaps you have already been asked for data from a previous project, but then simply did not have the time to look up the data again or provide it in time?

Rise your hand if you agree …

How many of you have several versions of the final report of a project …

```
.
..
DETER_FinalReport_final.doc
DETER_FinalReport_final2.doc
DETER_FinalReport_final2_mb.doc
DETER_FinalReport_final2_mb_ia.doc
DETER_FinalReport_final2_mb_pr.doc
DETER_FinalReport_final2_mb_pr_final.doc
DETER_FinalReport_final2_mb_pr_final2.doc
DETER_FinalReport_final2_mb_pr_final3.doc
```

Rise your hand if you agree …

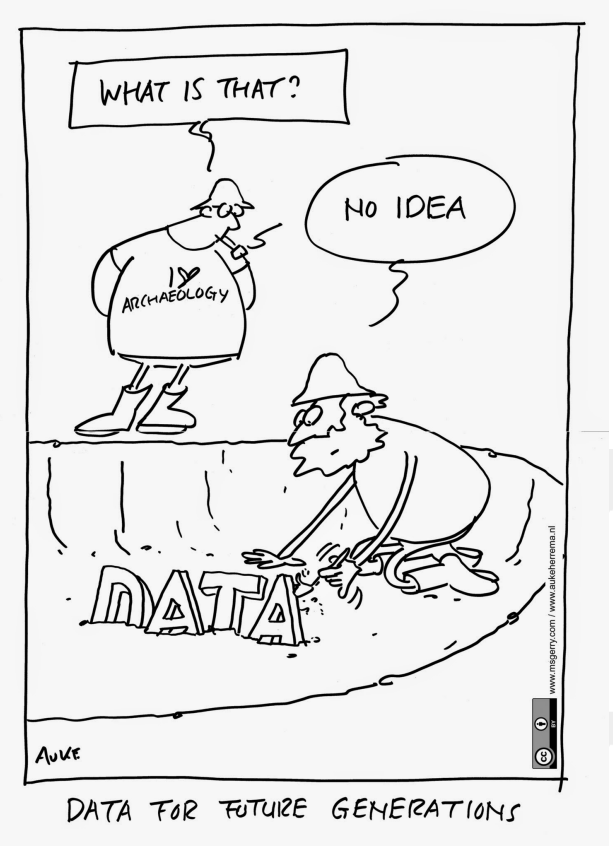
… and don't even really have the final version?

Research Data Management Services, current Workflows and DMPs

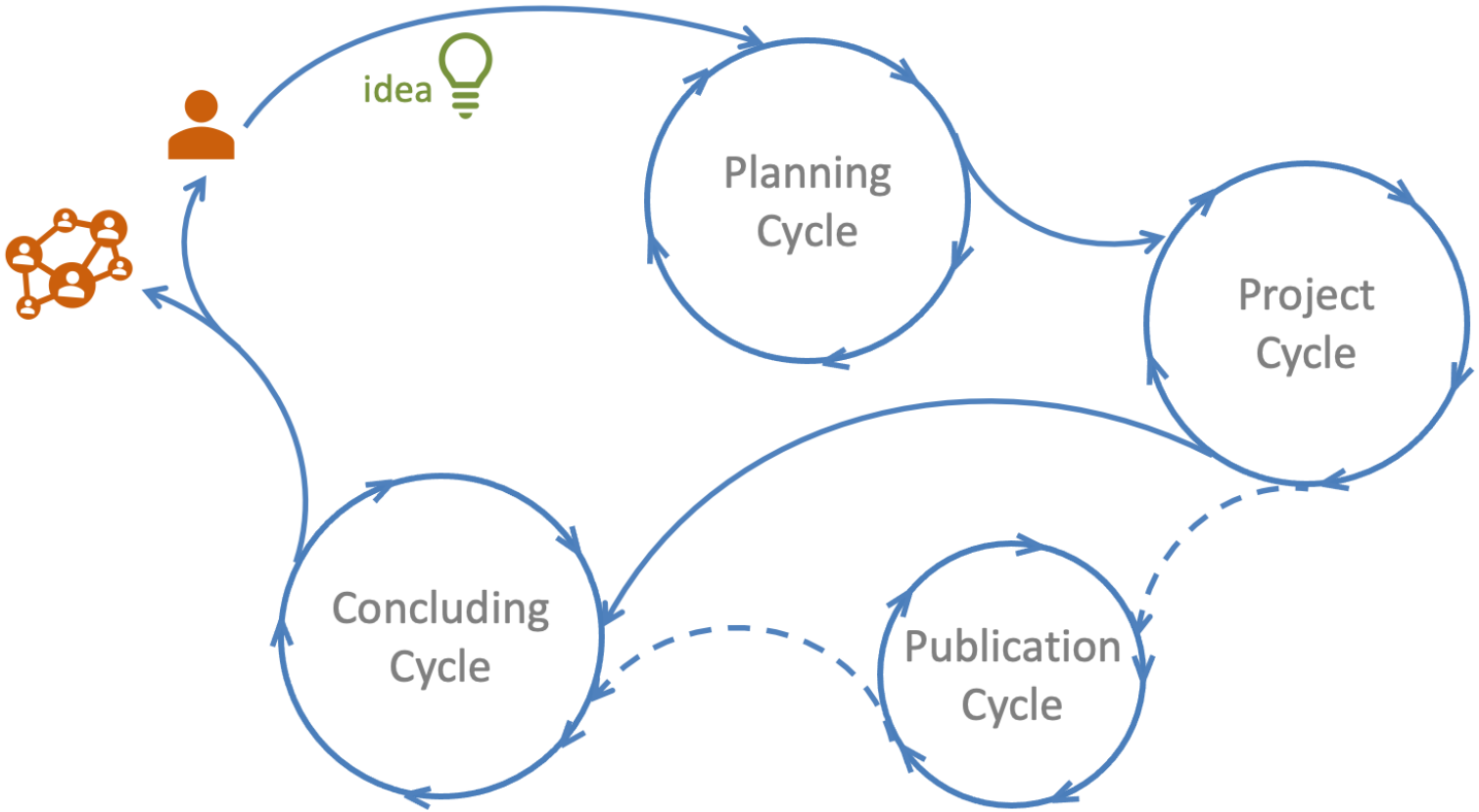
DKRZ USER WORKSHOP

13. 10. 2022

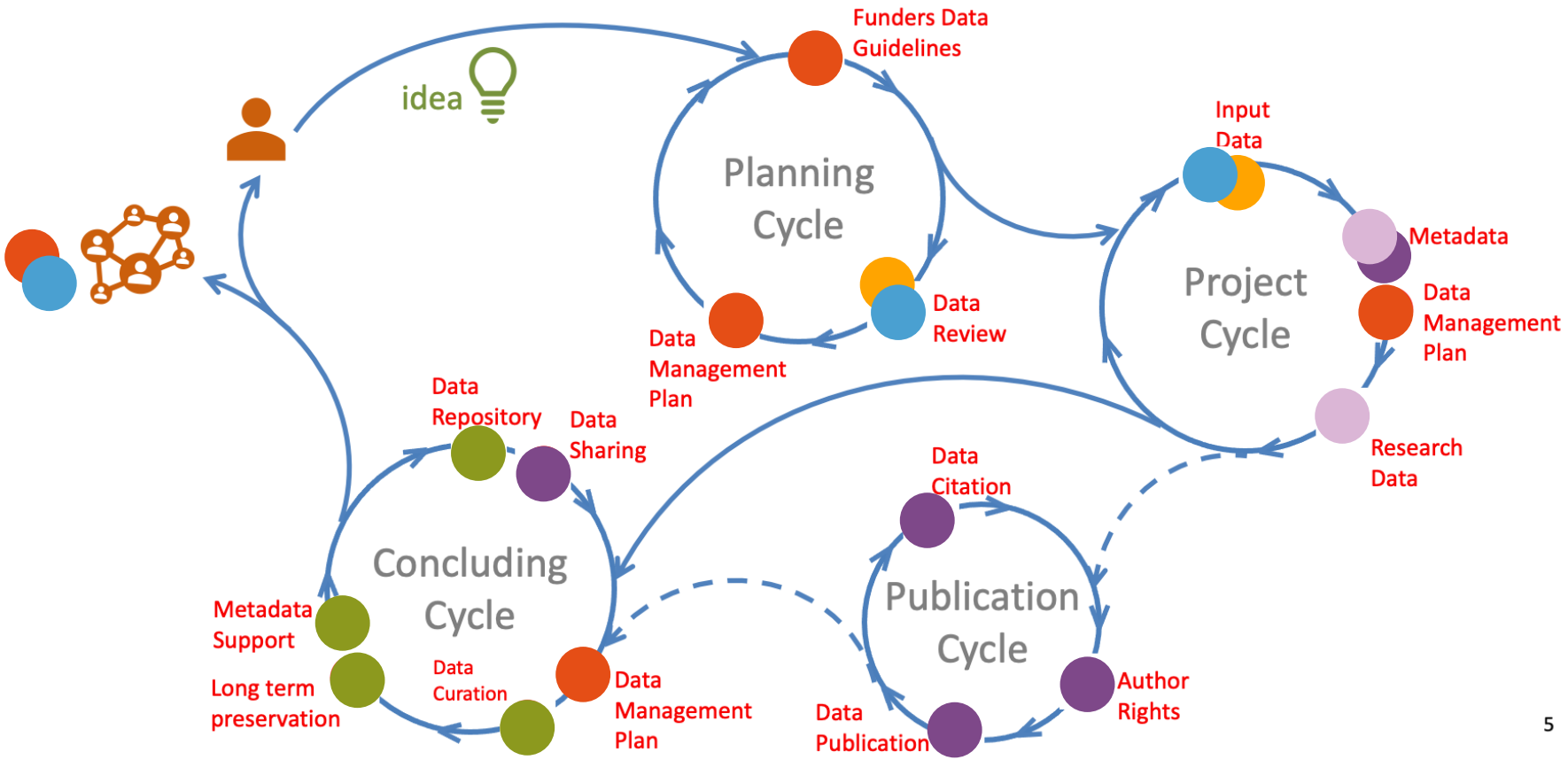
Ivonne Anders, Karsten Peters-von Gehlen
& ALL@DM



The Project Cycle

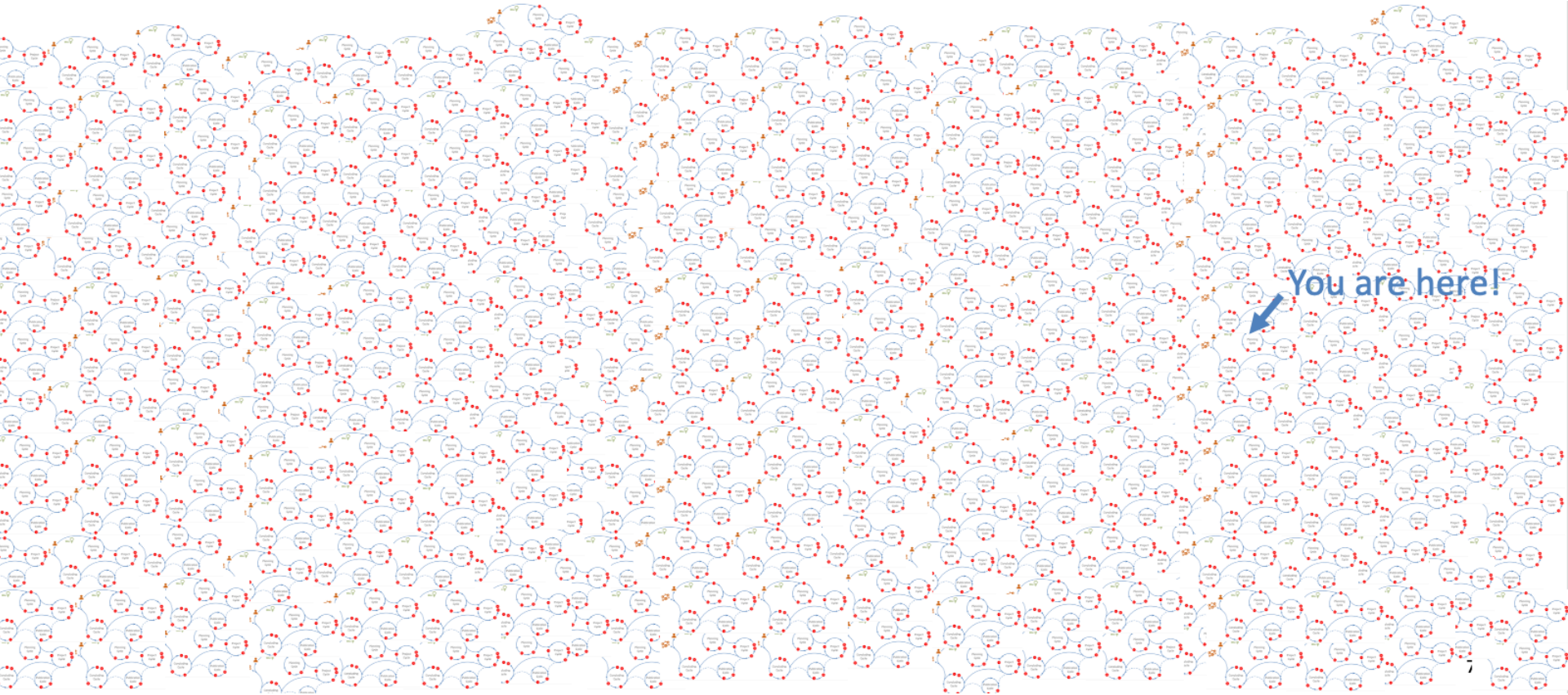


Data – the most faithful companion



5

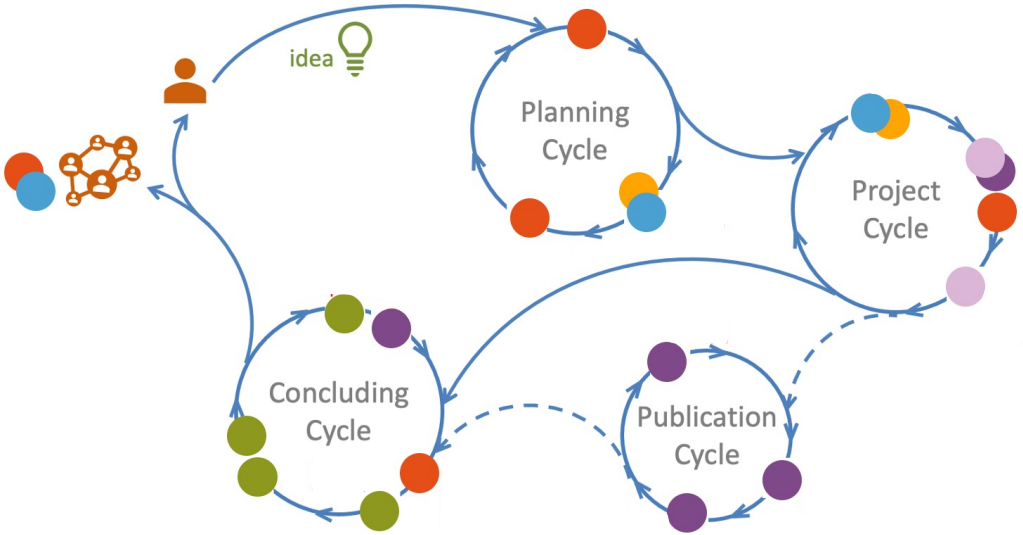
Data at DKRZ



Challenges

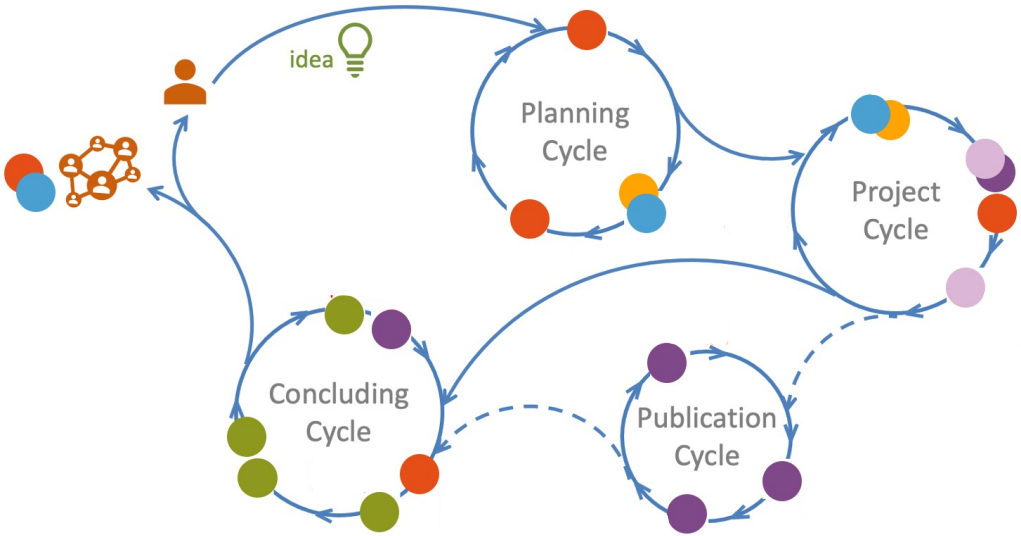
… current developments in Earth System Modeling demand for efficient workflow strategies to cope with ever increasing data amounts, complexity of analyses, server-side processing and distributed databases

Project Cycle



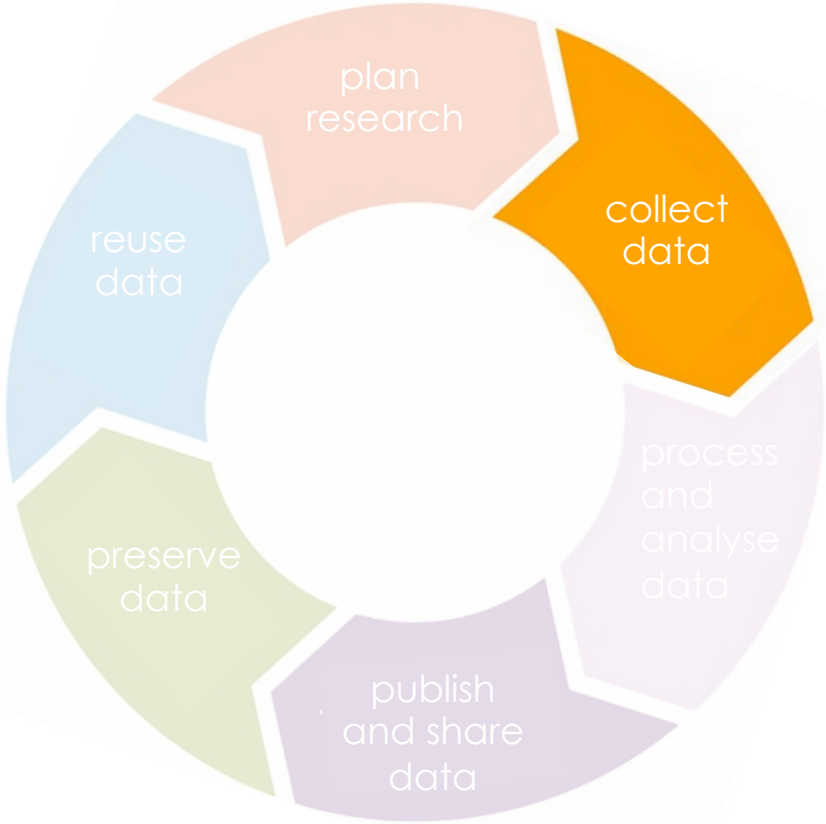
RDM Services

Project Cycle



Data Life Cycle

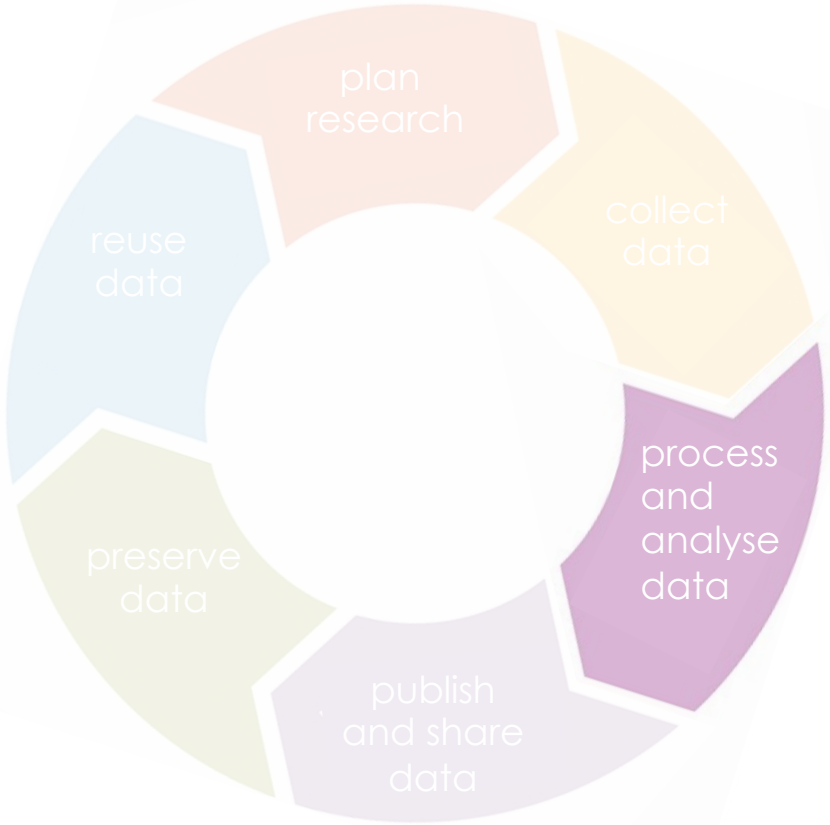
RDM Services



- Model simulation output
  - Everything related to Earth System Science
- Data collections hosted by DKRZ
  - CMIP3/5/6
  - CORDEX
  - ERA5, ERA-Interim
  - …
- Archived data
  - Project-based tape archive
  - World Data Center for Climate (WDCC)
  - DOKU

https://www.wdc-climate.de/

RDM Services



- Discipline-specific and centrally maintained software stack
  - Dedicated support available
  - Expanded if required
- Intake catalogs
  - CMIP5/6
  - CORDEX
  - ERA5
  - DYAMOND
  - nextGEMS

INTAKE

- Experimental STAC implementation

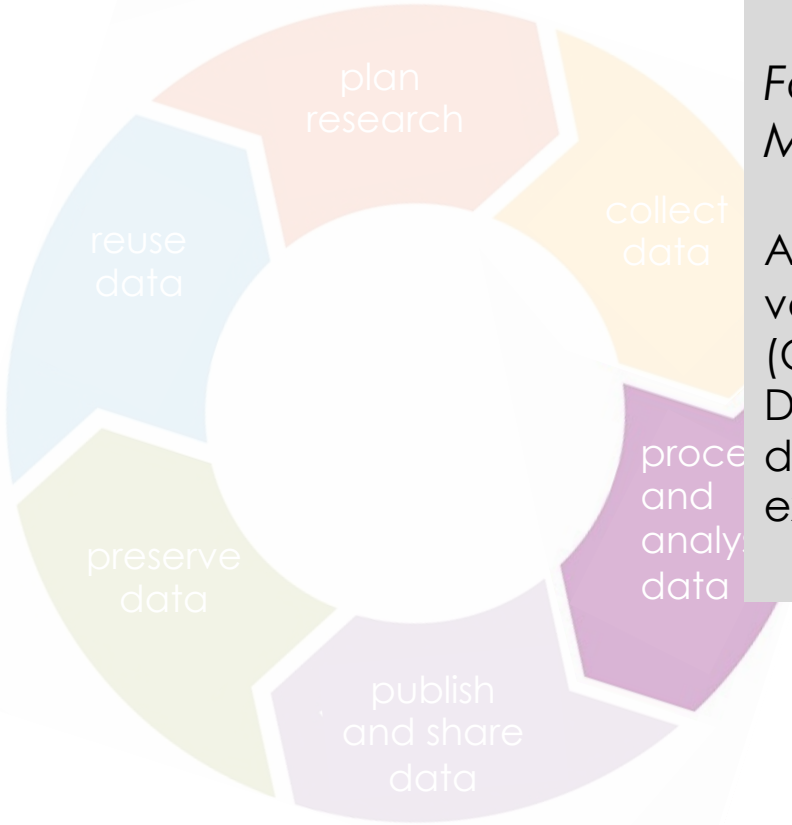
STAC

- Jupyterhub @DKRZ
  - Direct access to HPC resources (compute and disk)
    https://jupyterhub.dkrz.de/

jupyterhub

- Data standardization support
  - Community- or project-specific (depending on user requirements)

RDM Services



Fabian Wachsmann and Marco Kulüke

Availability and accessibility of large-volume datasets hosted at DKRZ (CMIP3/5/6, ERA5, CORDEX, DYAMOND,…) using metadata-driven data access (catalogs), including an extensive hands-on session.

INTAKE

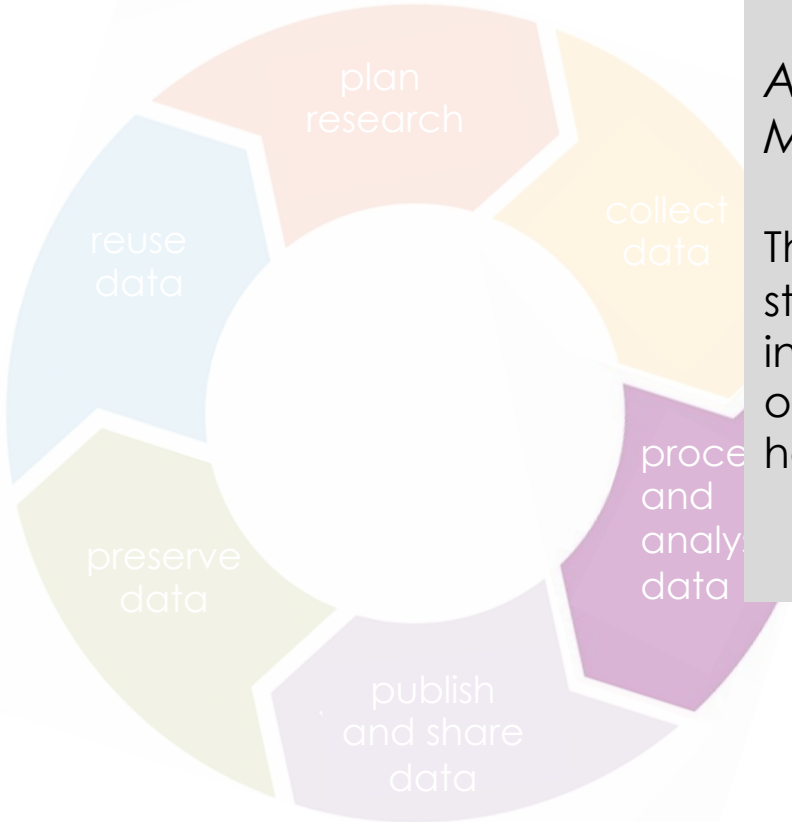
STAC

- Direct access to HPC resources (compute and disk)
  https://jupyterhub.dkrz.de/

jupyterhub

- Data standardization support
  - Community- or project-specific (depending on user requirements)

RDM Services



Angelika Heil and Martin Schupfner

The basics and benefits of data standardization including an introduction of tools and services offered at DKRZ + an extensive hands-on session

INTAKE

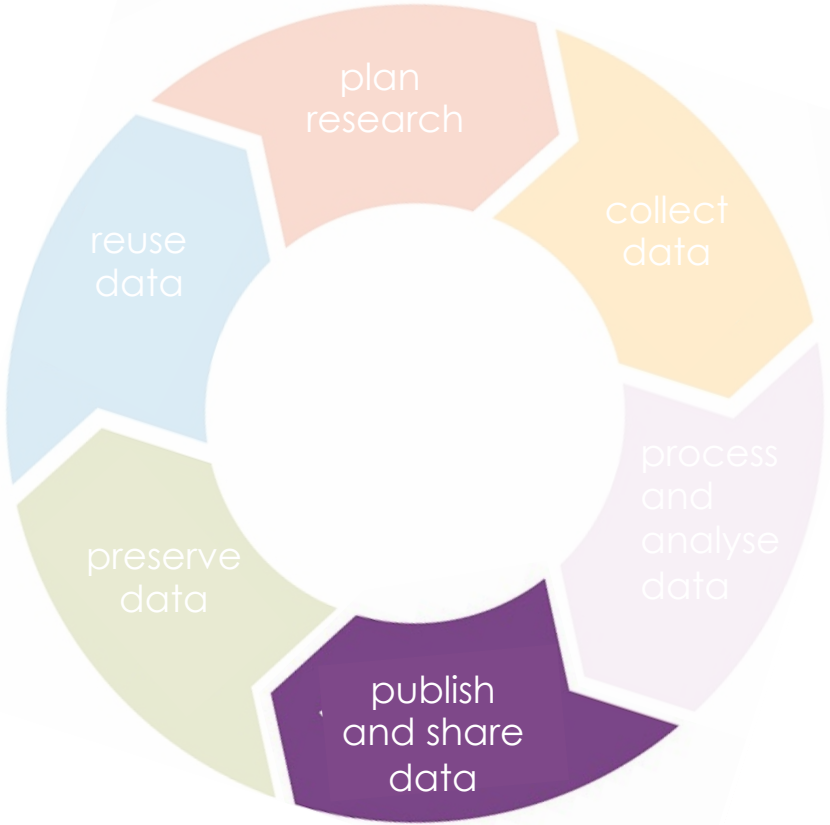
STAC

- Direct access to HPC resources (compute and disk)
  https://jupyterhub.dkrz.de/

jupyterhub

- Data standardization support
  - Community- or project-specific (depending on user requirements)

RDM Services



- Sharing via DKRZ cloud (swift object store, 1.5PB current capacity) https://swiftbrowser.dkrz.de/
- Publication in WDCC or DOKU (long-term archival services)
  - DataCite DOI for WDCC

  https://www.wdc-climate.de/
- Dissemination via ESGF

  https://esgf-data.dkrz.de/
- Jupyter-Notebooks

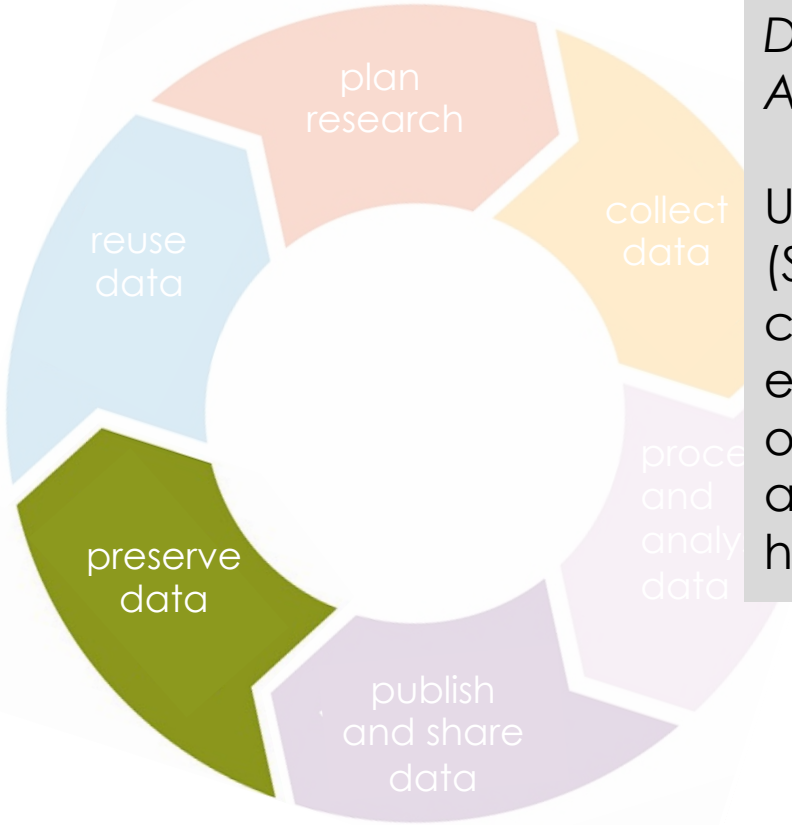
  https://jupyterhub.dkrz.de/
- Project-related support for organizing data publication/sharing/dissemination

RDM Services

- Tape archive for everyday use
  - More than 300 PB available
  - Metadata-driven search and access
- DOKU
  - Archival of project-related reference data
  - PIDs
  - Available to DKRZ users
- WDCC (World Data Center for Climate)
  - CoreTrustSeal certified domain-specific repository
  - FAIR compliant (Peters-v. G. et al., 2022)
  - DOIs available
  - Available to the global community

https://www.wdc-climate.de/

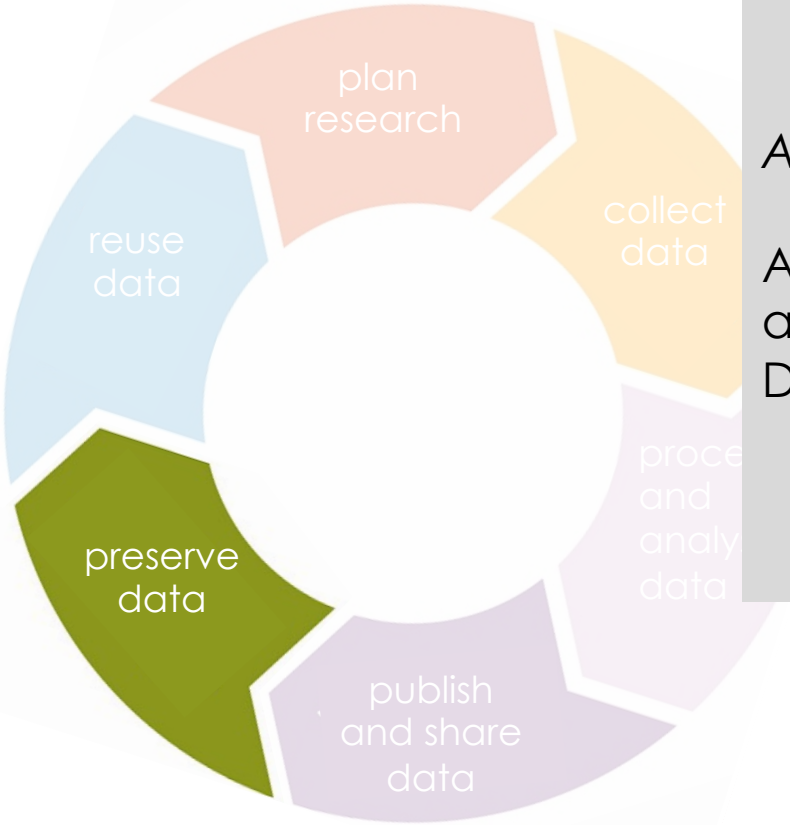
RDM Services



Daniel Heydebreck and Andrej Fast

Using the DKRZ tape archive (StrongLink) and its associated command-line tool slk in your everyday workflow with a focus on its metadata-enabled data access framework; also including hands-on

RDM Services



Andrea Lammert

An overview of DKRZ long-term archiving services WDCC and DOKU

STRONG LINK

access

erence data

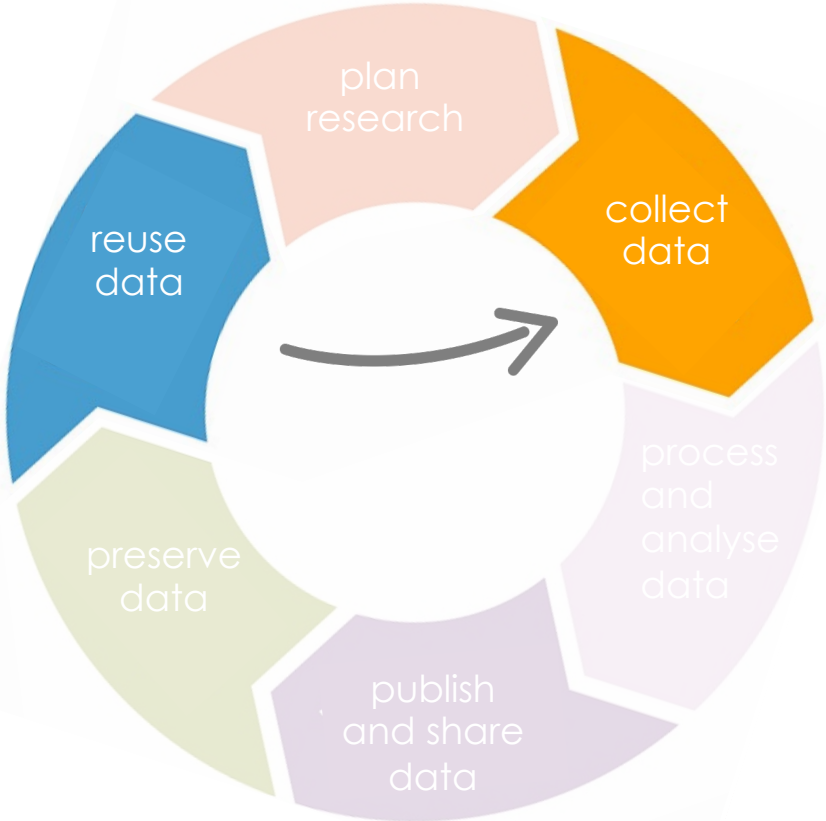
Climate)
n-specific

- FAIR compliant (Peters-v. O. et al., 2022)
- DOIs available
- Available to the global community

WDC CLIMATE
https://www.wdc-climate.de/

Findable Accessible Interoperable Reusable

RDM Services



Created by 4B Icons from the Noun Project

www.digitalbevaring.dk

RDM Services

We always recommend a good planning!

See a Data Management Plan as your "Project's Brain" and a therefore a living document.

Created by Teewara soontom from Noun Project

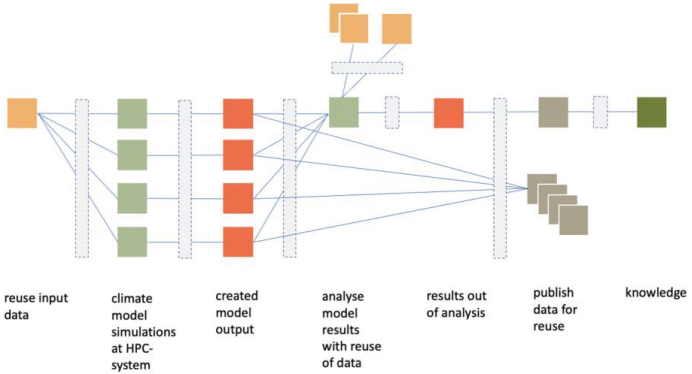
We support e.g. large projects. Get in contact with us!

Comming back to our challenges

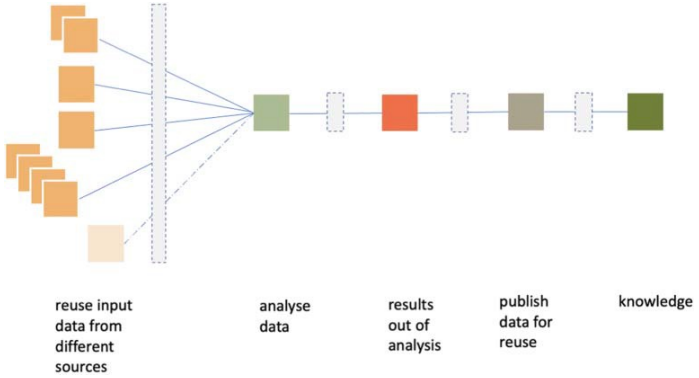
… current developments in Earth System Modeling demand for efficient workflow strategies to cope with ever increasing data amounts, complexity of analyses, server-side processing and distributed databases

Analysed Workflows

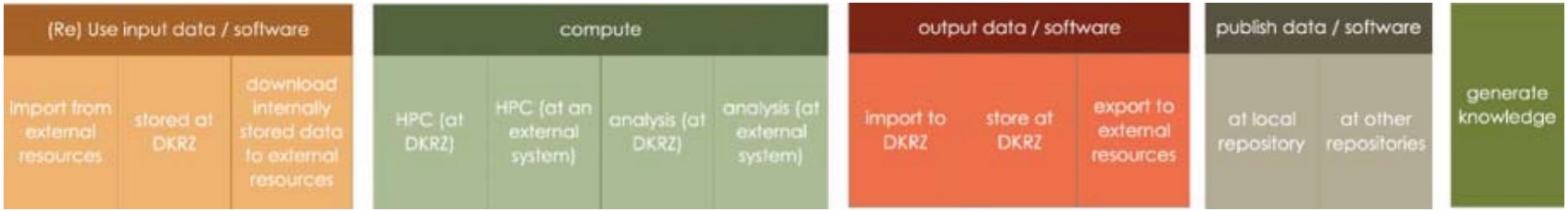
The modeller



The data analyst / reuser



Workflow elements



Anders et al., 2022

a wishlist …

"efficient access to and discovery and processing of (very) large in-house and remote datasets"

"end-to-end ESM workflow solution including automated output data handling"

Keep model output routines untouched

Publication-ready data package in-line with FAIR principles

Cataloguing of performed ESM simulation setups to enable efficient reuse of large datasets for different scientific questions

Automated setup of model runs given research question (or reuse of existing data)

Online diagnostics and data compression

Possibility of very specific database queries

Comprehensive metadata cataloguing and metadata-driven access across storage tiers

Interfaces coupled directly to model output stream to ingest data in catalogued database

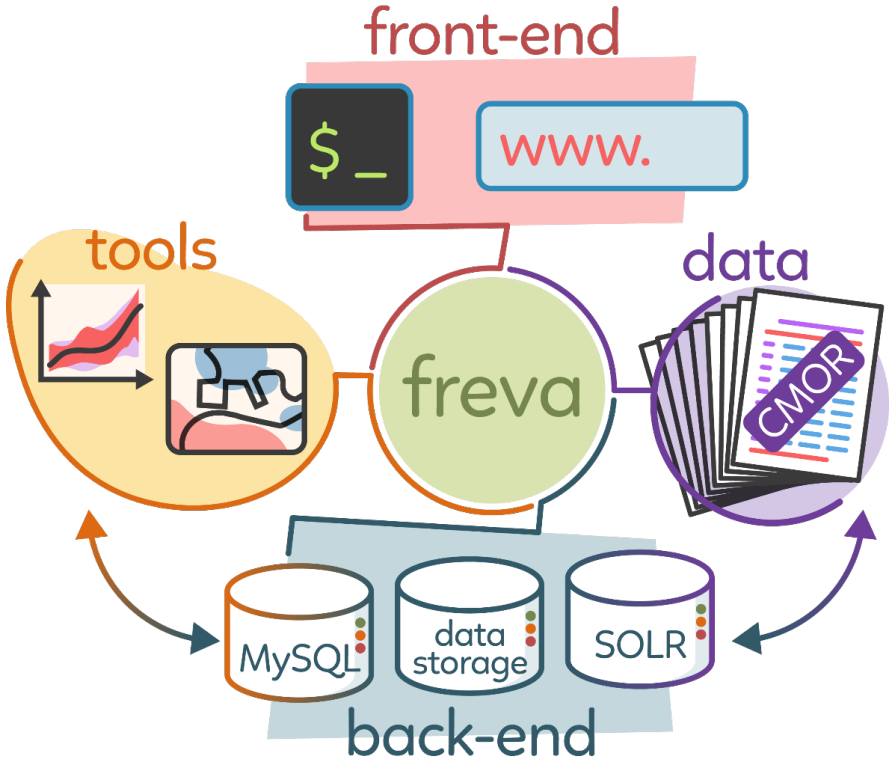
Stock of available and reusable processing scripts/workflows

Data formats allowing for fast access and processing

freva

Software infrastructure for standardized ***data and tool solutions*** *in Earth system science. Freva runs on* ***high performance computers, it*** *to handle* ***customizable*** *evaluation systems of research projects, institutes or universities.*

See Kadow et al. (2021) or DKRZ Documentation for more details



Freva - Framework

freva

Software infrastructure for standa
data and tool solutions *in Earth sy*
science. Freva runs on ***high perfo***
computers *to handle* ***customizab***
evaluation systems of research pr
institutes or universities.

See Kadow et al. (2021) or D
Documentation for more deta

Martin Bergemann and
Etor E. Lucio Eceiza

Get your hands on Freva, the Free Evaluation System for Earth System Modeling, and explore the possibilities available for your workflow of the future.