Research Data Management in PalMod

PalMod DMP Workshop

11. Feb. 2020

Minutes of the Discussion Session

Karsten Peters (DKRZ)

Discussion

Topics (so far):

- Software publication/sharing/versioning?
- Intra-project workflows/timelines?
- Managing storage load during project runtime
- Organisation of data publication via ESGF/WDCC/GFZ/ PANGAEA/…
- Collaboration based on distributed datasets?
- Presentation/Results presented at the kick-off meeting
- Computing resource proposal

Software publication/sharing/versioning?

- Software is part of scientific data
- Access to model versions of e.g. Palmod phase I?
- **Licenses**
- Further discussion is needed
- Clarify points of contact for software as well

intra-project workflows (how to organise)?

- CC.2 needs phase 1 products
- Compile a preliminary list of published products from phase 1 (lists of weblinks, document, publications, directory on mistral, …)
- Need for variables (look at PMIP proposed variables), what is the desired time frequency
- Status of data for internal use; very detailed list to keep updated
  - Google spreadsheet? GEOMAR project webpage?
  - Keep it efficient/lean
- Communication
  - There will soon be a PalMod email list beginning of March 2020
  - Wiki on the PalMod homepages
  - Introduce the existing GEOMAR system again as “operational” system earlier than the kick-off (Kerstin, Hela); Mattermost instance already installed

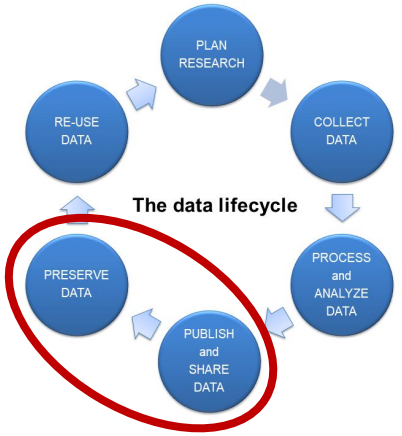
Managing storage load during project runtime

- Handling of “surprise simulations”
  - Support from DKRZ needed? **Automatic** and **catalogued** entry into the tape archive?
  - Part of the esm-tools, Hendryk installed it
  - New HSM-system coming next year will have more intelligent technical capabilities to handle e.g. metadata for intelligent data handling
- Need to discuss variable output before the runs; distinguish between essential variables and “nice to have”
- Changing responsibilities for data storage provision from the provider to the user
- Keep in mind the PMIP variables and decide on them
- Compile a first list with needed variables from the data users as soon as possible

Data Publishing and Sharing

„Publishing data“ happens at any point where data are actively made available for third parties to access

A multitude of services exists for **Earth System Science data publishing and sharing**, either out- or in-house at DKRZ….



DKRZ Swiftbrowser

[DKRZ-cloud](#) for easy sharing of (large-volume) datasets with anybody, e.g. RCEMIP.

Metadata creation is up to the user, access through sharing links, no back-up

ESGF Earth System Grid Federation

[Global infrastructure for disseminating large datasets](#), e.g. CMIP, MPI-Grand Ensemble (by the end of the year)

Fileformat (netCDF) and file organization has to strictly follow given standards, usually no back-up

DKRZ LTA DOKU

DKRZ‘s lightweight long-term archiving service for **project or publication reference data**

Metadata creation is up to the user, but minimal set is required; any data format; findable in CERA web-GUI AND via pftp (/hpss/doku) from mistral

PANGAEA.

Certified data publisher for Earth & Environmental Data, **focusing on observations,** including the possibility of assigning DOI‘s

Extensive metadata required, technical quality control, any format, global findability

WDC CLIMATE

DKRZ‘s certified long-term archiving service **with a heritage of focusing on model data**, including the possibility to assign DOI‘s to datasets

Extensive metadata required, technical quality control, GRIB/netCDF/geoTIFF file formats, global findability

Organisation of data publication (ESGF/WDCC/etc)

- PalMod has some resources to spend on hardware (about 1PB of disk space for ESGF)
- Swiftbrowser is not a secure persistent data hosting service
- Have to decide on variables to keep
- ESGF-published data have to be cmorized, but you can define your own project requirements and standards (which should still be very similar to CMIP)
- **Distinguish between everyday workflow and which variables to actually publish for external access**
- No defined experiment protocol yet (which helps to define variables and project standards)
  - Again, start from the PMIP tables
- Global attributes referring to the funder should be required
- Zuwendungsbescheid: “data should be published…”

Collaboration based on non-HH datasets

- Non-Hamburg datasets available on DKRZ disks
- If Hamburg should be considered a central location for large external datasets, this has to be decided
  - On a small-scale, the data should be centralised
- What is small- and what is large-scale?
- A feasible workflow for cross-HPC platform access to large datasets should be thought of