Workshop on Meta-data (MD) Generation During Experiments

21/22 January 2014

At Deutsches Klimarechenzentrum, Hamburg, contact: Toussaint@dkrz.de

Web: https://verc.enes.org/ISENES2/events/workshop-on-meta-data-md-generation-during-experiments

Agenda (not as planned but close to what has happened):

Tue 21 Jan (room #023, ground floor): Workflows at the different sites

12:30 Arrival, Snacks for Lunch

13:00 Frank Toussaint	Welcome and Logistics
13:10 Reinhard Budich	Overview
13:20 Mark Elkington	Climate model workflows at MetOffice
13:40 Sébastien Denvil	Climate model workflows at IPSL
14:00 Alok Gupta	Climate model workflows at Uni Norway, Bergen

14:20 DISCUSSION

14:30 - 15:00 Coffee...

15:00 Stephanie Legutke	Meta Data Generation of ES Experiments in the DKRZ ESModelling environment
15:50 V.Balaji (remote)	Situation at GFDL

16:15 - 17:00 DISCUSSION:

What will the organisations do specifically for IS-ENES-2?

18:00 Walk to dinner

Wed 22 Jan (room #023, ground floor)

09:00 Kerstin Fieg & Luis Kornblüh - MD on provenance:
An attempt to collect provenance data during production

10:20 Mark A. Greenslade (remote) - CIM structure and MD input into it

10:45 DISCUSSION

10:45 - 11:15 Coffee...

11:15 Tobias Weigel PIDs for data and metadata

Tobias Weigel Further PID usage scenarios

Thereafter: DISCUSSION

12:00 - 13:00 Snacks for Lunch

13:30 Michael Böttinger Präsentation of Data Visualisation at DKRZ

Participants (in total):

V.Balaji (remote)
Reinhard Budich
Sébastien Denvil
Mark Elkington
Kerstin Fieg
Mark A. Greenslade (remote)
Alok Kumar Gupta
Stefan Kindermann
Luis Kornblüh
Stephanie Legutke
Hans Ramthun
Torsten Rathmann
Martina Stockhause
Frank Toussaint
Tobias Weigel
Heiner Widmann
Hans-Hermann Winter

Workshop Results

There were no estimates on how far any institute intends to get with automated metadata (MD) generation within the timeframe of IS-ENES2. There was, however, agreement that this is a necessary process and that MD generation should be linked closer to the data production than presently the case.

Discussions

At the *Workshop on Metadata Generation During Experiments*, discussions on various topics were held. Three groups of widely agreed statements on metadata (MD), quality checks (QC), and Persistent Identifiers (PID) were protocoled.

Metadata capture

- There was agreement that a good approach to keep MD is to have one or more DB to keep them.
- Two presently used forms of automated metadata generation were identified: In ESGF the MD are written to the file headers. In a second step ("data publication") the MD are collected, written to a database (DB) and made searchable. In other systems (e.g., DKRZ, GFDL) the filling operations of the DB continuously more or less reflect the process status and are finished when the data generation process is.
- DB aside or DB not aside: This is rather a social question than a technical one. Practical work shows that it often makes sense to keep the DB separated from the modelling environment and to split access rights and responsibilities. This often results in cleaner and clearer workflows and liabilities.
- Various DB (two or three per system) might be used as for MD of data description (classical MD), MD of the workflow (provenance etc.), and for further topics (model description, quality check results, etc.). Here, too, it probably does make sense to split responsibilities.

Recommendations on Quality Checks and its Metadata

- There should be well communicated consensus on the technical requirements to the data (formats, headers, vocabulary used, etc.).
- These agreements need to be published before the modelling centres start their work as they affect the output software.
- Tools that check on these agreements or on data content need a detailed description of what they really check. This yields especially for verification checks.
- Checks on data content should result in warnings, not in errors as right or wrong depends on the data usage.

- Similar to the errata information in CMIP5 there should be a means of informing users and data centres of deviations from the agreed technical requirements on formats, headers, etc, to assess possible additional efforts needed for the use of these data.
- It needs to be pointed out very clear that data which do not follow the agreed technical requirements might not be published and distributed by the data centres.

Discussion on Persistent Identifiers

- PID to mark CIM instances: In the CMIP5 project, the attribution of a model or ensemble description, mostly in CIM formatted xml files, to the data turned out to be rather difficult. The different components of the data storage key (Data Reference Syntax, DRS) were regarded as single attributes in the CIM. No common key had been agreed on. Here the assignment of PID to the CIM instances and to the data might be a solution.
  The flexibility of PIDs also will solve another problem: The granularity level of the CIM descriptions often differs from what is needed for data handling and distribution. Here PID Collections might help to adjust the degree of aggregation.
- PID to make versioning easier: A common issue is a correct handling of different versions of model output, as, e.g., for corrections often only some of the files of a model run need to be touched. Those operations should be documented, e.g. in a database. Here persistent PIDs can be of great help, as other pointers to the files might break after some time.
- It has, however, to be decided, whether a PID should be assigned to every single file or to an atomic dataset. In the former case PID Collections might help to achieve the latter.