Identifier Infrastructure Usage for Global Climate Reporting

IoT Week 2017, Geneva

Tobias Weigel
Deutsches Klimarechenzentrum (DKRZ)
World Data Center for Climate (WDCC)
CMIP Phase 6 (CMIP6)

Overview CMIP6 Experimental Design and Organization

The overview paper on the CMIP6 experimental design and organization has now been published in GMD (Eyring et al., 2016). This CMIP6 overview paper presents the background and rationale for the new structure of CMIP, provides a detailed description of the CMIP Diagnostic, Evaluation and Characterization of Klima (DECK) experiments and CMIP6 historical simulations, and includes a brief introduction to the 21 CMIP6-Endorsed MIPs.

A brief summary can be found in the following overview presentation (CMIP6FinalDesign_GMD_161109.pdf) and below. After a long and wide community consultation, a new and more federated structure has been put in place. It consists of three major elements:

1. a handful of common experiments, the DECK (Diagnostic, Evaluation and Characterization of Klima) and CMIP historical simulations (1850 – near-present) that will maintain continuity and help document basic characteristics of models across different phases of CMIP,
2. common standards, coordination, infrastructure and documentation that will facilitate the distribution of model outputs and the characterization of the model ensemble, and
3. an ensemble of CMIP-Endorsed Model Intercomparison Projects (MIPs) that will be specific to a particular phase of CMIP (now CMIP6) and that will build on the DECK and CMIP historical simulations to address a large range of specific questions and fill the scientific gaps of the previous CMIP phases.

https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6
Scientific driver: Global climate modelling

- Operational phase ca. 2017-2021+
- Community-driven, aligned with IPCC AR6
- Global data volume in order of 100-250 PB
  - full replication impossible!

The climate data life-cycle

DKRZ Projects

1. Data Management Plan
2. DKRZ Storage
3. ESGF Preparation
4. ESGF Services
5. LTA DOCU
6. LTA WDCC
7. DataCite Data Publication

ESGF
Research Data Environment

Production + Data Processing
Dissemination
DOI publication
Standard Archiving
Dissemination During Project
Long-Term Archiving

M. Lautenschlager
The Earth System Grid Federation

http://esgf.llnl.gov
http://esgf-data.dkrz.de

DKRZ technical infrastructure and ESGF

HPSS tape
- 190 Pbyte capacity

Lustre file system
- 54 PByte

“Mistral” HPC
- 3.6 Pflops
- ~100,000 cores

Compute/ storage cluster
- VM servers, database servers
- Openstack cloud storage

S. Kindermann
Making it scalable requires additional effort

The PID registration and metadata update tasks are pushed to a message queueing system facilitating high availability and scalability...

...and then processed asynchronously.

Buurman, Weigel, Juckes, Lautenschlager, Kindermann: Persistent Identifiers for CMIP6 in the Earth System Grid Federation, EGU 2016
Properties stored in Handle records for ESGF

<table>
<thead>
<tr>
<th>Files</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>URL</td>
</tr>
<tr>
<td>aggregation_level</td>
<td>aggregation_level</td>
</tr>
<tr>
<td>url_replica</td>
<td>replaced_by</td>
</tr>
<tr>
<td>tracking_ID</td>
<td>replaces</td>
</tr>
<tr>
<td>checksum</td>
<td>errata_IDs</td>
</tr>
<tr>
<td>is_part_of</td>
<td>has_parts</td>
</tr>
<tr>
<td>DRS_ID</td>
<td>DRS_ID</td>
</tr>
<tr>
<td>file_size</td>
<td></td>
</tr>
<tr>
<td>file_name</td>
<td></td>
</tr>
</tbody>
</table>

 Kernel Information Profiles
Why do we care? What is the long-term strategy?

Drivers
- Data volume, complexity, audience
- Compute: I/O

Induced change
- Data life-cycle model
- File/object management practice

Solution space
- Architectural layering
- Processing to the data: New services, cultural change
- Automation

Insight and integrity (provenance, QC)
The users’ reality...

```
[handle@handle svr_10876]$ ls
admpriv.bin    handle_keystore
admpriv.bin.old logs
admpub.bin    prikey.bin
admpub.bin.old pubkey.bin
config.dct    serverCertificate.pem
contactdata.dct serverCertificatePrivateKey.bin
delete_this_to_stop_server serverCertificatePrivateKey.pem
[handle@handle svr_10876]$  
```
Type-Triggered Automated Processing (T-TAP)

- netCDF-Files
- Collection
- `<Metadata>` (xml)
- ? (third-party input)

Agent

Processing service (WPS)

- script

output

- well-defined ways to publish it (automatically)
- possible repacking into a new collection

multiple types, e.g. netcdf, xml, linked data, text reports, PROV record

described in DTR
Data processing perspectives

- Climate data analytics service (for EOSC)
  - Cluster-based, 2 pilot implementations, 2018+

- Copernicus Climate Change Service (C3S)
  - coordinated by ECMWF, operational 2018+
  - WPS-based service ecosystem with multiple deployments
End users, developers, and automated processes deal with persistently identified, virtually aggregated digital objects, including collections which are overlays on multiple network services which in turn are overlays on existing or future information storage systems.

Global Digital Object Cloud (GDOC)

Identifier Infrastructure Usage for Global Climate Reporting

L. Lannom / DFIG

Tobias Weigel (DKRZ)

09.06.2017
GDOC and reusable data service components
Thank you for your attention.

weigel@dkrz.de