Versioning for CMIP6 in the Earth System Grid Federation

... and PIDs!

Data preparation

Initial registration

Version updates

End-user tools

References

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Cliparts from http://www.openclipart.org

Tobias Weigel, Katharina Berger, Stephan Kindermann, Michael Lautenschlager
German Climate Computing Center (DKRZ)

Introduction

Prototype impression
Motivation

- No common ESGF approach to versioning, unclear processes
- Demonstrate usefulness of wide-scale low-level PID usage within operational e-infrastructure
- Controlled versioning at this scale will be new for CMIP6
Weigel, Berger, Kindermann, Lautenschlager

Versioning for CMIP6 in the Earth System Grid Federation

Under development

ESGF publisher:
• publish
• unpublish
• update metadata

ESGF web GUI
(via cog)

ESGF user collection builder

ESGF DB

esgscan tool

populate

query/search

register

Handle API v8 (REST)
PIT API
Collections API

Handle System
Type Registry

.solr

Future options (>2016)

Generic / community-independent

End-user CLI tools

Stand-alone PID web tools

Automated verification service

query additional information (if available)

DataCite DOI assignment process

Collections API

Type Registry

Research Data Alliance (RDA)

CMOR

.nc

.nc

.nc

.pit

Create collection

query

read, check conformance

.read, check conformance

query additional
information (if available)

DataCite DOI assignment process

Collections API

Type Registry

Research Data Alliance (RDA)

CMOR

.nc

.nc

.nc

.pit

Create collection

query

read, check conformance

.read, check conformance

query additional
information (if available)

DataCite DOI assignment process

Collections API

Type Registry

Research Data Alliance (RDA)

CMOR

.nc

.nc

.nc

.pit

Create collection

query

read, check conformance

.read, check conformance

query additional
information (if available)

DataCite DOI assignment process

Collections API

Type Registry

Research Data Alliance (RDA)

CMOR

.nc

.nc

.nc

.pit

Create collection

query

read, check conformance

.read, check conformance

query additional
information (if available)

DataCite DOI assignment process

Collections API

Type Registry

Research Data Alliance (RDA)

CMOR

.nc

.nc

.nc

.pit

Create collection

query

read, check conformance

.read, check conformance

query additional
information (if available)

DataCite DOI assignment process

Collections API

Type Registry

Research Data Alliance (RDA)
What is required?

- Technical development (esgf publisher)
- Agreement on pioneering nodes
- Definition of policies to be enforced
- DKRZ Handle service and future coordination

Until end of 2015...
Essential versioning policies

- Versioning can only be trustworthy if everyone adheres to the policies.
- Enforce use of ESGF tools as opposed to unmonitored changes in the file system.
- Unified version numbers: YYYYMMDDxx
  - recommended for all future projects using ESGF
  - mandatory if automated version management is to be used
Prototype impression

Weigel, Berger, Kindermann, Lautenschlager

Versioning for CMIP6 in the Earth System Grid Federation

17.04.2015 - EGU2015-9445

Dataset PID

File PID

PID := prefix+tracking_id

What happens when clicking on a PID?
ESGF publisher:
- publish
- unpublish
- update metadata

ESGF web GUI (via cog)

ESGF user collection builder

Collections API
- register
- populate
- query/search
- create collection
- query additional information (if available)

ESGF DB
- query
- populate
- query/search

Handle API v8 (REST)
- Handle System
- Type Registry

Collections API
- create
- read, check conformance

PIT API

solr

CMOR

Data preparation

esgscan tool

Future options (>2016)

End-user CLI tools

Stand-alone PID web tools

Automated verification service

Generic / community-independent

Under development

DataCite DOI assignment process

Data preparation

CMOR

Handle System

Type Registry
Under development

ESGF publisher:
- publish
- unpublish
- update metadata

esgscan tool

ESGF DB

ESGF web GUI (via cog)

ESGF user collection builder

Stand-alone PID web tools

Future options (>2016)

End-user CLI tools

Automated verification service

Collections API

Handle System

Handle API v8 (REST)

CMOR

Data preparation

Raw files

Write PID in netcdf header

e.g. using CMOR; also determine version number

Configuration of concrete PID syntax according to common agreements (e.g. within CMIP6)

Register PIDs

Files visible

Register additional PIDs for aggregates

using PID tools to be provided by EUDAT

Add replica locations to PID records

It is also possible to let data owners add additional locations through a dedicated service (e.g. provided by EUDAT)

What’s in a PID record?

CMOR

Data preparation

Handle Server

Data Node provider
ESGF publishing process

Data provider / modelling center

Data provider / modelling center
Example PID records (DWD obs4MIPs prototype)

Dataset: 10876/ESGF/4ee9d37b-6454-44bf-b3ef-e738b2ecedb4

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td><a href="http://bmbf-ipcc-ar5.dkrz.de/thredds/esgcet/3/obs4MIPs.FUB-DWD.SSMI-MERIS.mon.v20140616.html">http://bmbf-ipcc-ar5.dkrz.de/thredds/esgcet/3/obs4MIPs.FUB-DWD.SSMI-MERIS.mon.v20140616.html</a></td>
</tr>
<tr>
<td>DRS name</td>
<td>obs4MIPs/observations/FUB-DWD/Obs-SSMI-MERIS/obs/mon/atmos/prw</td>
</tr>
<tr>
<td>Publication date</td>
<td>2014-06-16</td>
</tr>
<tr>
<td>Version number</td>
<td>20140616</td>
</tr>
<tr>
<td>Children</td>
<td>[&quot;10876/ESGF/a9b1bfbc-4b73-4295-8ed6-6b586bf1be02&quot;, ... ]</td>
</tr>
</tbody>
</table>

File: 10876/ESGF/a9b1bfbc-4b73-4295-8ed6-6b586bf1be02

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRS name</td>
<td>prwErr_SSMI-MERIS_L3_v1-00_200301-200812.nc</td>
</tr>
<tr>
<td>Publication date</td>
<td>2014-06-16</td>
</tr>
<tr>
<td>Checksum (MD5)</td>
<td>F49ee38e24e819b5d04c534f6ed7b375</td>
</tr>
<tr>
<td>Size</td>
<td>50989760</td>
</tr>
<tr>
<td>Parent</td>
<td>10876/ESGF/4ee9d37b-6454-44bf-b3ef-e738b2ecedb4</td>
</tr>
</tbody>
</table>
Under development

ESGF publisher:
- publish
- unpublish
- update metadata

ESGF web GUI (via cog)

ESGF user collection builder

Collections API
- Collections API
- PIT API
- Handle API v8 (REST)

Type Registry
- Handle System

Handle API v8 (REST)

PIT API

Handle System

CMOR

esgscan tool

ESGF DB

generate

query/search

query

read, check conformance

create collection

DataCite DOI assignment process

Stand-alone PID web tools

End-user CLI tools

Automated verification service

Future options (>2016)

Generic / community-independent

Initial registration

register

query additional information (if available)
Under development

ESGF publisher:
- publish
- unpublish
- update metadata

Esgscan tool

ESGF DB
ESGF web GUI (via cog)
ESGF user collection builder

Stand-alone PID web tools

Future options (>2016)
- End-user CLI tools
- Automated verification service

Collections API
- Handle
- System
- PIT API
- Handle API v8 (REST)
- Type
- Registry
- register
- populate
- query/search
- query
- create
- collection
- read, check
- conformance
- query
- additional information (if available)

Raw files
- Write PID in netcdf header
e.g. using CMOR; also determine version number

Configuration of concrete PID syntax according to common agreements (e.g. within CMIP6)

What’s in a PID record?
- CMOR
- Write PID in netcdf header
- Register PIDs
- Files visible

Register additional PIDs for aggregates
using PID tools to be provided by EUDAT

Add replica locations to PID records
It is also possible to let data owners add additional locations through a dedicated service (e.g. provided by EUDAT)

Data provider / modelling center

Data Node provider

ESGF publishing process

Initial registration
- register

Handle API v8 (REST)
Handle System

Home
On updates, the initial publication process is largely repeated, but the publisher detects the existing files and arranges old and new files in a collection accordingly.
Parts of this process should be supported by EUDAT/ESGF tools to make it more scalable and reduce current manual effort.
Basic PID resolution

hd:10876/ESGF-2b8e6aef-3806-44d9-9eda-d10d3cbefce

Aggregate? Singleton? Tombstone?

singleton

File PID

aggregate

Dataset PID

tombstone

Individual information page services provided by ESGF based on EUDAT tools

Web landing pages could offer: data download, versioning information, replication information, ...
Possible command line tools:
- wget for PID‘ed data with smooth authentication
- info tool
- get latest version
... all Python-based!

Possible web tools:
- Generic viewer across communities (PIT use case)
- Provenance tracing tool
Envisioned EUDAT2 PID services architecture

- Verification tools and services
- B2* Services
- EUDAT PID service (epicclient.py)
- PID system base services (CRUD, distribution)
- Advanced PID services (viewer, reverse lookup, queueing system, ...)
- Mass management tools
- Collection service (lapis / Collection WG)
- Operational tools (monitoring, siteinfo, ...)
- HSv8 native REST
- Solr indexing servlet
- Relational DB (*SQL)
- Apache solr
- Handle System 8 with embedded Jetty

Future EPIC service concept?
(focus however on organizational aspects/QA)


http://www.handle.net
http://cnri.reston.va.us