Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

THE IDEA
Integrate data-related information into the federated data infrastructure ESGF (Earth System Grid Federation).

THE PRESENT SITUATION
Data and different kinds of additional data-related information are collected in separate and poorly connected repositories.

Within climate research these are:
- ESGF for data and use metadata
- ES-DOC for metadata on experiments and simulations
- IPSL’s Errata Module for errata and version information
- WDCC’s Citation Service for data citation information
- CharmE’s annotation module

EXTERNAL METADATA REGISTERED IN ESGF
Core Metadata is registered in ESGF, making the development of the external component dependent on the ESGF development.
Core metadata include a reference to detailed information (URL, PID).
This approach is suitable for data-related information available on file or dataset granularity and collected jointly with the data.

EXTERNAL METADATA REFERENCED FROM ESGF
The development of the external component is independent of the ESGF development.
An API is provided and referenced from the ESGF.
This approach is suitable for data-related information available on a different granularity and collected independent of the data.

THE IDEA
Integrate data-related information into the federated data infrastructure ESGF (Earth System Grid Federation).

THE PRESENT SITUATION
Data and different kinds of additional data-related information are collected in separate and poorly connected repositories.

Within climate research these are:
- ESGF for data and use metadata
- ES-DOC for metadata on experiments and simulations
- IPSL’s Errata Module for errata and version information
- WDCC’s Citation Service for data citation information
- CharmE’s annotation module

EXTERNAL METADATA REGISTERED IN ESGF
Core Metadata is registered in ESGF, making the development of the external component dependent on the ESGF development.
Core metadata include a reference to detailed information (URL, PID).
This approach is suitable for data-related information available on file or dataset granularity and collected jointly with the data.

EXTERNAL METADATA REFERENCED FROM ESGF
The development of the external component is independent of the ESGF development.
An API is provided and referenced from the ESGF.
This approach is suitable for data-related information available on a different granularity and collected independent of the data.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

THE PRESENT SITUATION

Data and different kinds of metadata are collected separately. Users and archive managers have to use multiple interfaces to gather information.

Data and different kinds of additional data-related information are collected in separate and poorly connected repositories.

Within climate research these are:

- ESGF for data and use metadata
- ES-DOC for metadata on experiments and simulations
- IPSL’s Errata Module for errata and version information
- WDCC’s Citation Service for data citation information
- CharmE’s annotation module

Data Preparation

CMOR

Data Ingest

ESGF data node

ESGF Publication

ESGF index node

Metadata Ingest

External repository

Metadata Publication

Long-Term Archival

Stable data enriched with metadata

EXTERNAL METADATA

EXTERNAL METADATA

EXTERNAL METADATA
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

**Data Preparation**
- CMOR

**Metadata Ingest**
- External repository

**Data Ingest**
- ESGF data node

**ESGF Publication**
- ESGF index node

**Metadata Repository**
- Collects and publishes information independent of ESGF.

**REGISTERED METADATA**
- e.g., errata

**REFERENCED METADATA**
- e.g., citation

**EXTERNAL METADATA**

**API as CONNECTION to ESGF provided by metadata repository.**

**Long-Term Archival**
- Stable data enriched with metadata

**EXTERNAL METADATA REFERENCED FROM ESGF**

- The development of the external component is independent of the ESGF development.
- An API is provided and referenced from the ESGF.
- This approach is suitable for data-related information available on a different granularity and collected independent of the data.

---

EGU2015-8404

REFERENCE METADATA - CITATION TYPE -

Metadata repository collects and publishes information independent of ESGF.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

Data Preparation

CMOR

Metadata Ingest

External repository

Core Metadata

Registration

ESGF Publication

ESGF data node

ESGF data node

Metadata Ingest

External repository

Metadata Ingest

External repository

Core Metadata Registration

External Metadata

REGISTERED METADATA - ERRATA TYPE -
Metadata repository registers core information including links to detailed information in ESGF.

ESGF provides registration mechanism for core metadata of external repository.

REGISTERED METADATA (e.g. errata)

DATA

Data Ingest

ESGF data node

ESGF Publication

ESGF index node

Core Metadata

Long-Term Archival

Stable data enriched with metadata

EXTERNAL METADATA
REGISTERED IN ESGF
Core Metadata is registered in ESGF, making the development of the external component dependent on the ESGF development.
Core metadata include a reference to detailed information (URL, PID).
This approach is suitable for data-related information available on file or dataset granularity and collected jointly with the data.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

FUTURE SITUATION
Data and metadata can be searched for, accessed and viewed together, hiding the infrastructure's complexity.

DATA
- Data Preparation
  - CMOR

REGISTERED METADATA
(e.g. errata)
- Metadata Ingest
  - External repository

REFERENCED METADATA
(e.g. citation)
- Metadata Ingest
  - External repository

ESGF Publication
- ESGF data node
- ESGF index node

Core Metadata Registration

Long-Term Archival
Stable data enriched with metadata
Data Preparation
CMOR

Data Ingest
ESGF data node

ESGF Publication
ESGF index node

DATA INFRASTRUCTURE

ESGF is a system for the management, access and analysis of climate data.

ESGF’s architecture is based on a system of autonomous and distributed nodes:

- Data is stored at multiple nodes, and served through local data and metadata services.
- Nodes exchange information about their data holdings and services.

ARCHITECTURE

ESGF's architecture is based on a system of autonomous and distributed nodes:

- Data is stored at multiple nodes, and served through local data and metadata services.
- Nodes exchange information about their data holdings and services.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

ESGF Quality Control Working Team (ESGF-QCWT)

ESGF-QCWT aims to integrate valuable external information into the ESGF.

These additional documentations:
• are collected asynchronous and independent of the data, and
• are usually provided on different granularities.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

DATA

Metadata Ingest
External repository

REGISTERED METADATA
(e.g. errata)

Data Preparation
CMOR

Data Ingest
ESGF data node

ESGF Publication
ESGF index node

Core Metadata Registration

REFERENCED METADATA
(e.g. citation)

Metadata Ingest
External repository

Metadata Publication

Long-Term Archival
Stable data enriched with metadata

2 TYPES of INTEGRATION

• REFERENCED METADATA – CITATION TYPE: completely independent development / loosely connected
• REGISTERED METADATA – ERRATA TYPE: core metadata registered in ESGF

USAGE of INTEGRATION TYPES

REFERENCED METADATA:
• central repository for referenced metadata
• workflow of metadata collection independent of data preparation
• independent development of external metadata repository → easy to implement in ESGF

REGISTERED METADATA:
• several decentralized repositories for registered metadata
• combined workflow for metadata and data preparation
• coordinated development of external repositories and ESGF → effort to implement in ESGF
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, gilpsl@ipsl.jussieu.fr)

Data Preparation
- CMOR

Metadata Ingest
- External repository

Data Ingest
- ESGF data node

ESGF Publication
- ESGF index node

Long-Term Archival
- Stable data enriched with metadata

REGISTERED METADATA
- (e.g. errata)

REFERENCED METADATA
- (e.g. citation)

DATA

Metadata repository collects and publishes information independent of ESGF.

CITATION EXAMPLE

1. Citation information on simulation datasets collected in parallel to the data collection.

2. Citation repository provides API as stable interface / ESGF include API call in portals.

3. LTA uses API to add citation information (or asks ESGF index for the link to citation information).
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

DOI data publication as final step in a 3 level Quality Control process. A data citation includes all datasets belonging to a simulation. Data citation information is available at WDCC and IPCC DDC.

WDCC’s CITATION Service for CMIP5 by tracking_id

More services of CMIP5 quality and citation repository available at:
DOI data publication as final step in a 3 level Quality Control process. A data citation includes all datasets belonging to a simulation. Data citation information available at WDCC and IPCC DDC.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

CITATION within CMIP5
Citations of CMIP5 data are hidden and nearly invisible in the ESGF data portal. Moreover citations are partly wrong or outdated.

Quality Control Report is outdated (2013-11-25): QC level 2 is displayed instead of QC level 3. Therefore DataCite DOI and citation information are missing (see previous page).
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

CITATION within CMIP5
Citation services available are at quality and citation repository WDCC, but are not integrated into ESGF portals.

CITATION for CMIP6
Citations for CMIP6 data will be accessible in ESGF data portals.

Add “Data Citation” link in ESGF portals to access external citation service at WDCC.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

CITATION for CMIP6

At the citation repository the following components have to be developed:

1. Database schema to store preliminary citation information prior to LTA
2. GUI to insert / update citation information
3. API to access citation information for an ESGF dataset
4. Display citation information on html page

1.+2. Collect Citation information

I. Information on persons, institutes, titles provided

II. Create lists of persons and institutes

III. Select data collection and connect it to title, creator list (authors), and different contributor lists, e.g. Funder, ResearchGroup, …

IV. Automatic check of citation against project agreements before publication
3.+4. Access Citation information

At the citation repository the following components have to be developed:

1. Database schema to store preliminary citation information prior to LTA
2. GUI to insert / update citation information
3. API to access citation information for an ESGF dataset
4. Display citation information on html page

REQUEST for citation by controlled names (DRS)

REQUEST by standard names, controlled by project

DISPLAY option

DATA option

Creators
Title
Publication Year
Publisher
PID
Contributors...

RESPONSE displayed (html) or as data (e.g. json format)
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

CITATION for CMIP6
Citations for CMIP6 data are directly accessible in ESGF data portals.
Additional data options for machine access are provided.

DATA

Data Preparation
CMOR

Data Ingest
ESGF data node

ESGF Publication
ESGF index node

DATA INFORMATION

Metadata Ingest
External repository

Long-Term Archival
Stable data enriched with metadata

CITATION INFORMATION

Ingest in Citation repository

Metadata Publication

“Data Citation”
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, gilips@ipsl.jussieu.fr)

DATA

- Data Preparation
- CMOR

- Data Ingest
- ESGF data node

- ESGF Publication
- ESGF index node

ERRATA INFORMATION

- Metadata Ingest
- External repository

- Core Metadata Registration

CITATION INFORMATION

- Metadata Ingest
- External repository

- Metadata Publication

Long-Term Archival

Stable data enriched with metadata

REFERENCED METADATA – ERRATA TYPE –

Data publishers use errata to summarize production errors and their corrections.

SCATTERED ERRATA

General CMIP5 issues:

IPSL-CM5 issues:

CESM1 and CCSM4 issues:
http://www.cesm.ucar.edu/CMIP5/errata/

CNRM-CM issues:

…etc.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, glipsl@ipsl.jussieu.fr)

REFERENCED METADATA
– ERRATA TYPE –
Proof-of-concept:
IPSL errata module for CMIP5 data

REQUEST:
filename(s) + version(s) or tracking ID(s)

PHP MODULE + DATABASE

ANSWER:
Modifications/corrections history
Links to corresponding issues

IPSL-CM5 Errata

The IPSL database counts 140630 files and 23932 files have been modified/corrected (i.e., 17%).

Please upload a text file (see template) giving the tracking ID(s) or the filename and the version of the NetCDF file(s) for which you want information (i.e., versions, hierarchical modifications or corrections). This interface will only deal with original CMIP5 files from IPSL-CM5.

Text file: [Choose file] Aucun fichier choisi. [Send] (Max size = 2Mb)

(i) Search by tracking ID(s) is recommended. Tracking ID(s) can be found in global attributes of your file(s) using the NCO commands (e.g., “ Netcdump -0 ”).
(ii) Search by filename requires the dataset version. The version corresponds to the publication date and can be found in dataset properties from ESGF portal.

Please click on an errata number in the following table to get more details about the issues raised in the IPSL database (see color legend below):

<table>
<thead>
<tr>
<th>Short description</th>
<th>Affected file(s)</th>
<th>New version(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 Wrong soil depth axis from ORCHIDEE</td>
<td>sstf_Ti.map.nc</td>
<td>20120430 0</td>
</tr>
<tr>
<td>#2 Wrong pH computing</td>
<td>ph_Ti.map.nc</td>
<td>20120439 0</td>
</tr>
<tr>
<td>#3 Duplicated grid-cells on the edges of ocean grid</td>
<td>sstf_Ti.map.nc</td>
<td>20120439 0</td>
</tr>
</tbody>
</table>

• List and document all known errors and file versions in IPSL data,
• Simplify the search of annotations for one or a set of files.
Integration of external metadata into the Earth System Grid Federation (ESGF)

K. Berger¹, G. Levavasseur², M. Stockhause¹, and M. Lautenschlager¹

¹German Climate Computing Center (DKRZ, stockhause@dkrz.de), ²Institute Pierre Laplace (IPSL, giipsl@ipsl.jussieu.fr)

REFERENCED METADATA
– ERRATA TYPE –
Implementation to ESGF for a CMIP6 centralized errata

P2P

Dataset

File 1

File 2

© EGU2015-8404