ES-DOC for CMIP6 and ES-DOC Services: ESGF-ERRATA

Atif Ben Nasser
Guillaume Levavasseur
Mark Greenslade
Sébastien Denvil
**ES-DOC for CMIP6 status**

**Different collect process than CMIP5's:**
- About half of the documents (experiments, simulations, ensembles,...) **automated** (following ESGF publishing)
- The remaining (model, conformance to protocol, forcings, responsible party,...) produced by groups when ready – joined together via the “further_info_URL” attribute
- Multiple tools to create these documents (python library or notebooks, questionnaire,...).

**Ready for community review (Dec 2016):**
- Documentation work-flow for CMIP6,
- Type of information to be collected,
- WIP white paper describing the above.

**Currently in internal review:**
- Document creation tools: automated and UIs (py-esdoc, questionnaire, cdf2cim,...)
- Ocean, atmosphere, sea-ice and top level realms

**Working on:**
- Forcings description (with Tim Johns et al. e.g. IPCC Table 12.1) – timeline: Nov 2016
- Short model tables for papers (draft for ocean available) – Jan 2017
- Update science contents of other realms (with the community/WGCM) – Feb 2017
• Project to document CMIP6 well under-way.
• Building on CMIP5 experience (both good and bad !)
• Clear set of use cases
• Community review formalised (internal, WIP/WGCM, wider)
• More user friendly for groups:
  • Large fraction is automated
  • Starting model description from CMIP5 version
  • Beta testing for a period of 5 months (Oct 2016 – Feb 2017) with various actors (UKMO, GFDL, IPSL)
    • Possibility of adding two new groups (suggestions ?)
  • Documentation for all steps (+ overview as WIP white paper)
• Full scheduled community release: March 2017

• Looking ahead (posts CMIP6) to include other « realms »:
  • Regional models, downscaling
  • Evaluation & metrics, obs4MIP
ES-DOC and Errata Service

• Ensures data quality by providing issue status tracking.

• Relies on the PID Handle service for retracing past and future versions of datasets and/or files.

• Divided into two major pieces: Remote server and associated webservice and local client/Front-end.

• Currently in alpha phase, heading towards a community beta.

• Full community release: March 2017

• Prospective:
  • Exposing API to other services (such as ESGF CoG front-end and Synda) to ensure real time feedback on data status.
  • Incorporating the issue declaration process in the conventional publishing workflow, technically or through enforcement.
• Issues are stored within the PID service handles using Unique Identifiers (uid):

Each handle has a new attribute on the dataset level identifying the issue id (set to none if there's no issue with the dataset in hands).

• Errata service queries the PID service and proceeds to extract its issue id:
Using the PIDs genealogy tree structure, finding whether the queried version of dataset/file is safe to use or is affected by a running issue.

• Alpha release (Dec 2016) includes:
  • Errata Web-service suite and the related issue inspection algorithm.
  • Errata Service Front-End with search/filters features.
  • Errata-client enabling user interaction (data providers especially).

• Currently in internal review/ Working on:
  • Authorization and security issues:
    • Delegated to 3rd party: Github Oauth2.0
    • Implemented a security policy currently in internal review
ERRATA CRAWLER WORKFLOW:
Expectations:

Expected output
ERRATA CRAWLER WORKFLOW:
Expectations:

- **Available input:**
  - Dataset/File PID
  - Dataset/File id and version number
  - List of PIDs

- **Expected output:**
  - Data issue history
  - For file or dataset
  - List of datasets/files history

- **Constraints:**
  - Needs to be scalable
  - Tolerable complexity
Dataset A: {u'GDR_NUMBER': '20010101', ..., u'DRS_ID': 'cmip5.output1.MPI-M.MPI-DUMMY.atef.test.dataset.ABCD', u'FIXED_CONTENT': 'TRUE', u'REPLACED_BY': 'hdl:21.14100/37043d8e-ac5e-3843-a019-c03017cc68aa', u'AGGREGATION_LEVEL': 'DATASET', ..., u'HAS_PARTS': 'hdl:21.14100/d9053480-0e0d-11e6-a148-3e1d05defe78;hrl:21.14100/63fa73be-0e10-11e6-a148-4r1d05defe78;hrl:21.14100/28ju73be-0e10-11e6-a148-a7751ce7ec0c'}

Dataset B: {u'GDR_NUMBER': '20020101', u'REPLACES': '21.14100/AAE01BA2-8436-378D-84ED-5A06B9FBEEE46', ..., u'DRS_ID': 'cmip5.output1.MPI-M.MPI-DUMMY.atef.test.dataset.ABCD', u'REPLACED_BY': 'hdl:21.14100/e0560a9d-2227-3175-b943-fc26c427a923', u'AGGREGATION_LEVEL': 'DATASET', ..., u'HAS_PARTS': 'hdl:21.14100/d9053480-0e0d-11e6-a148-3e1d05defe78;hrl:21.14100/2a1d100e-0e13-11e6-a148-3e1d05koki66;hrl:21.14100/28ju73be-0e10-11e6-a148-a7751ce7ec0c'}

File.nc: {u'CHECKSUM_METHOD': 'SHA256', u'IS_PART_OF': 'hdl:21.14100/aae01ba2-8436-378D-84ED-5a06b9fbee46;hrl:21.14100/37043d8e-ac5e-3843-a019-c03017cc68aa', ..., u'FILE_NAME': 'atef_esgf_testfile_temperature.nc', u'FILE_VERSION': 'fv1', ..., u'FILE_SIZE': '1291743184', u'CHECKSUM': 'fbab91863fcc67cf118d698c0ac210f79c6e7118a3f3c585f311c0d5d36cacf2', u'AGGREGATION_LEVEL': 'FILE', ...}
ES-DOC Errata Service:

ERRATA CRAWLER WORKFLOW:

• A file handle has no information about the next version of the file or the preceding version

• 3rd part logical tree libraries were considered to reconstruct the tree and easily extract the errata ids, such as NetworkX for python...
ERRATA CRAWLER
ES-DOC Errata Service:
ERRATA CRAWLER WORKFLOW:

Dataset_Record#v1
+ replaced_by: str
+ replaces: str
+ id: str
+ errata_id: str

Dataset_Record#v2
+ replaced_by: str
+ replaces: str
+ id: str
+ errata_id: str

Dataset_Record#v3
+ replaced_by: str
+ replaces: str
+ id: str
+ errata_id: str

First appearance of file in current version

Last appearance of file in current version
ES-DOC Errata Service:
ERRATA CRAWLER WORKFLOW

• Medium complexity, scales up to million files/datasets resolution with little trouble (theoretical worse case scenario $O(n^2)$)

• Average cyclomatic complexity and great maintainability index:

```
[root@pc-296 errata]# radon cc handle_service/harvest.py
handle_service/harvest.py
  F 21:0 harvest_errata_information - A
[root@pc-296 errata]# radon cc handle_service/crawler.py
handle_service/crawler.py
  F 17:0 crawler - C
```

Radon Cyclomatic Complexity

```
[root@pc-296 errata]# radon mi handle_service/crawler.py
handle_service/crawler.py - A
```

Radon Maintainability Index
ES-DOC Errata Service
Errata Crawler limits, perspectives

• Inherits sequential behaviour from PID handle server.
• Straightforward for datasets, but not for files.
• Open room for improvements, according to community needs and expectations, update the WIP paper accordingly.
ES-DOC Errata Service
Errata Crawler limits, perspectives


• Github client repository: https://github.com/ES-DOC/esdoc-errata-client

ES-DOC Errata Service

✓Ask away...