Agreement on data management and ESGF publication workflow

G. Levavasseur[1], A. Stephens[2], A. Iwi[2], K. Berger[3], N. Carenton[1], A. Bennasser[1], S. Denvil[1] and S. Ames[4]

[1]Institute Pierre Simon Laplace (IPSL), [2]Science and Technology Facilities Council (STFC), [3]German Climate Computing Center (DKRZ), [4]Lawrence Livermore National Laboratory (LLNL)

Introduction

The ESGF publication workflow depends strongly on the data management of each datanode. Consequently, the high flexibility of the publication command-line allows partner institutes to build their own publication workflows according to their local data policies. Unfortunately, without common use of the publication tools the ESGF archive became difficult to use and manage, especially for projects containing thousands of datasets (e.g., CMIP5). To ensure a high data quality for CMIP6, the IS-ENES Data Task Force is investigating around the ESGF publication workflow, taking into account as many use cases of existing data management from ESGF partners as possible.

II. Publication stakeholders - Who does what?

The ESGF publication workflow includes many steps that require the skills of different actors. A fine division of the required tasks allows each institute to organize its resources:

The **Scientific Data Provider**
only deals with the climate data (i.e., production, issues, etc.).

The **Quality Manager**
ensures the readiness of climate data for publication (i.e., controlled vocabulary, data quality, etc.).

The **Data Manager**
stores the formatted climate data on the file system with the appropriate directory tree and versioning.

The **Data Node Manager**
is in charge of publication actions through the ESGF publisher.

The **End User**
requests the ESGF front-ends and downloads the climate data.

III. Key recommendations/enforcements

Versioning check-up

- ☑ A publication units cannot be published with two different versions anywhere in ESGF (i.e., master + replicas).

Local file system versioning

- ☑ The directory tree has to follow the *Data Reference Syntax* (DRS) of the corresponding project.
- ☑ New version string has to be in agreed format.
- ☑ A review of `drslib` library is recommended to manage directory contents between versions using symlinks.

Mapfile = key by-product of the publication process

- ☑ The mapfile has to embed the version number(s) and the SHA256 file checksum.
- ☑ The use of mapfiles is preferred as input to the publisher.
- ☑ One mapfile per publication unit allows to set our own "mapfile-granularity" and control our publications.

Node configuration

- ☑ Move out the `project` sections of the `esg.ini` into `esg.<project_id>.ini` files.
- ☑ `esg.<project_id>.ini` files aims to check the DRS and its controlled vocabulary.
- ☑ Provide `esg.<project_id>.ini` via GitHub with the appropriate options list and maptables.

Notifications

- ☑ Notify all above actors that the dataset has been published.
- ☑ Notify the interested end users of the new version, version retraction or removal.

IV. Publication test suite

Publication involves many different tasks that can be performed in different sequences. One way to counter the inherent unpredictability of the publication process is to develop a test suite that can be run by any node manager to check that a set of prescribed tasks generates a consistent outcome.
(cf. "*ESGF Publication 2015 Sprint" report*)

V. Interactions with other services

The Handle Service

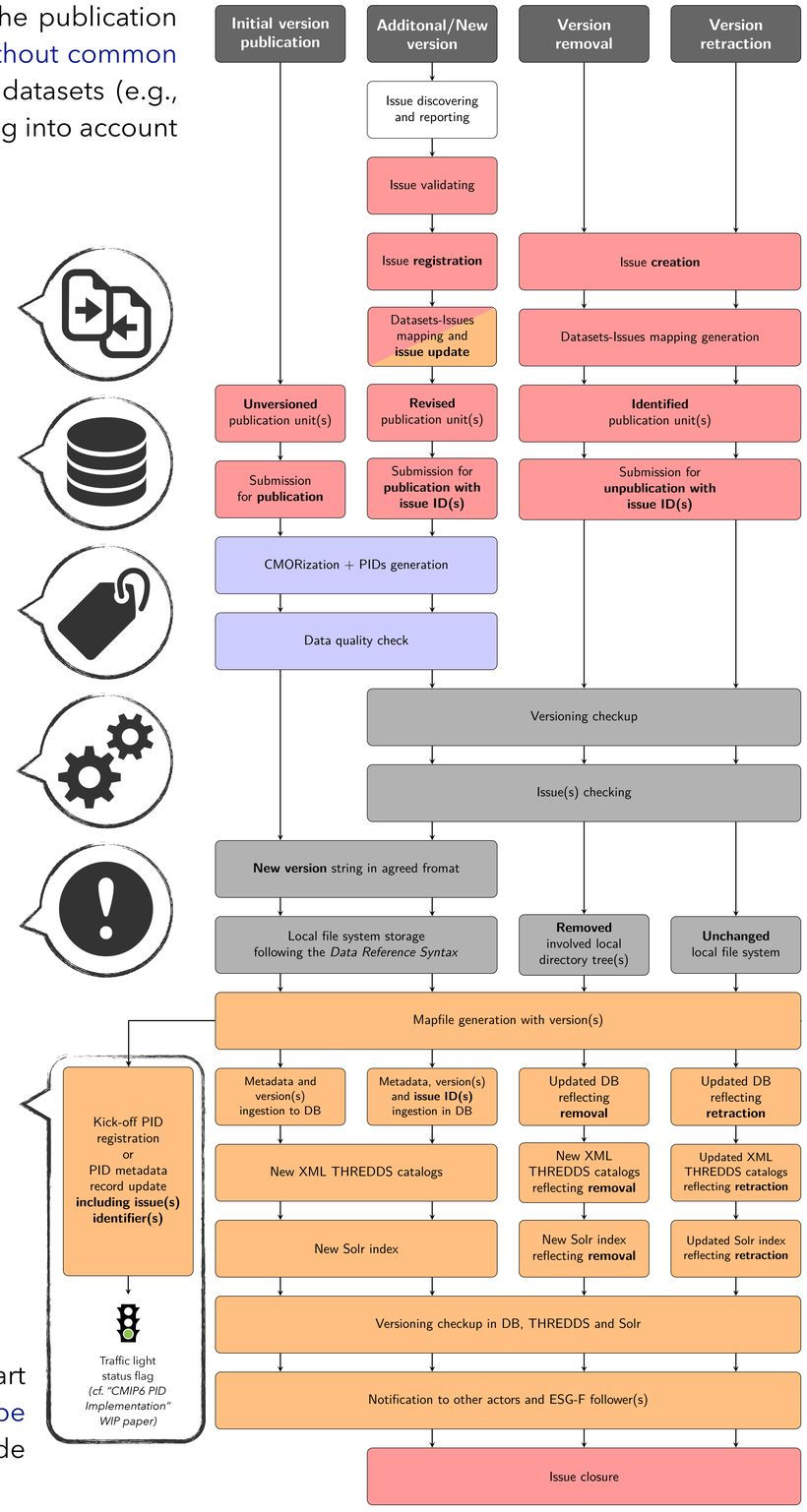
The Persistent IDentifier system records any change as a "filiation" or "two-way" links between two dataset versions. To remove or retract a dataset version, the corresponding PIDs are updated at the Handle Service tracking any unpublished data. A Solr instance will be serve PID information as part of the HS, to support requests as:

- ☑ If a dataset has previously been published and find out about previous versions,
- ☑ If a file is already registered,
- ☑ If two files (i.e., different checksums) are sharing the same PID.

The Errata Service

(cf. *ESGF F2F 2015 "CMIP6 Errata" poster*)

I. Recommended publication workflows



Conclusion

In order to improve the end-user experience through new ESGF services (e.g., PID, errata, etc.), the IS-ENES Data Task Force aims to start the groundwork for best practice in publishing our CMIP6 data as soon as possible. All recommendations and enforcements that will be implemented in the publisher's code can be discussed to be compliant with most of the ESGF partners (e.g., modelling groups, node managers, institution, etc.).