Un service web errata pour CMIP6

Guillaume LEVAVASSEUR
Atef BENNASSER
Mark GREENSLADE
Sébastien DENVIL

Sciences de l'environnement
Institut Pierre Simon Laplace

19 Novembre 2018

Plan:

- Introduction
  - Pourquoi un service errata ?
  - Le Service Errata ESGF en bref
- Démo (1/2) : naviguer dans l'errata sans ramer
- Comment identifier mes données ?
  - Des données aux multiples "facettes"
  - Vous avez dit PID ? Kesako ?
- Démo (2/2) : Une analyse sans "fautes"
- Conclusion

Agence Nationale de la Recherche ANR - CONVERGENCE MODÈLES NUMÉRIQUES
is-enes
cnrs
SORBONNE UNIVERSITÉ

Pourquoi un service errata pour CMIP6?

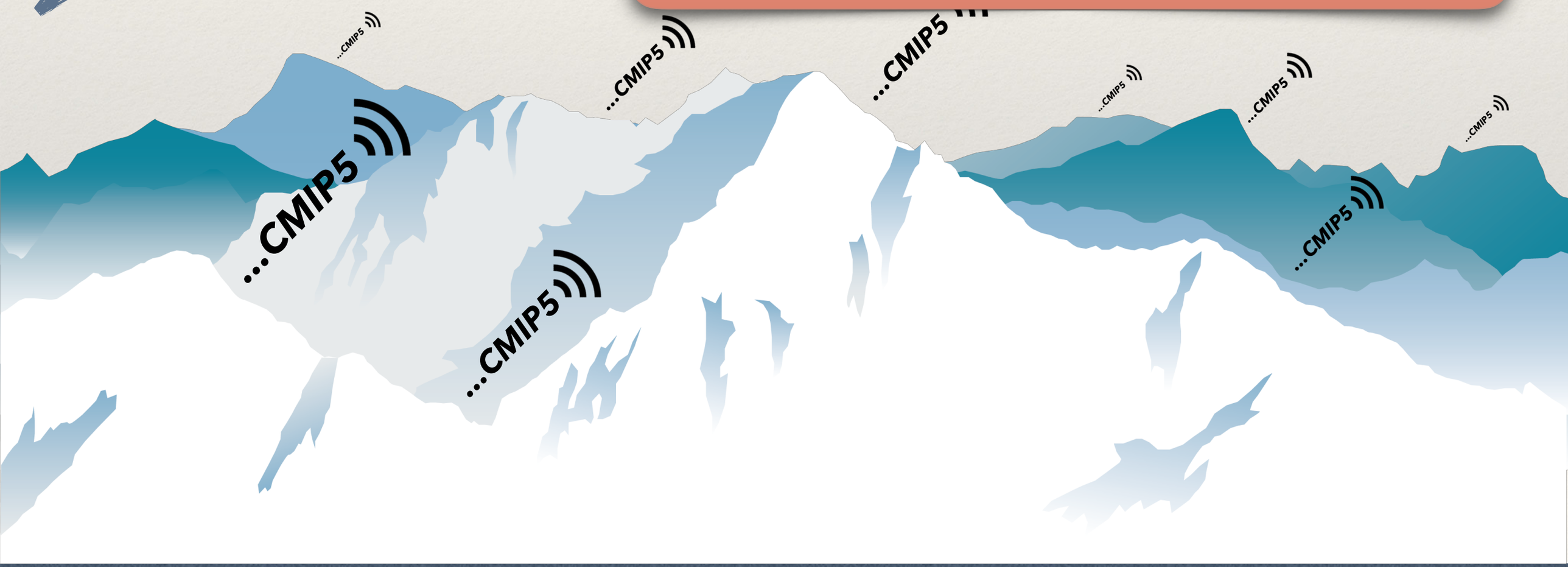
Allô Houston ?
J'ai un truc bizarre avec
un fichier CMIP5?

Pourquoi un service errata pour CMIP6?

Allô Houston ?
J'ai un truc bizarre avec un fichier CMIP5?

Objectifs:

- **Centralisation** : centraliser les erreurs rencontrés sur les données d'un même projet,
- **Standardisation** : documenter les erreurs selon un même schéma,
- **Traçabilité** : enregistrer et suivre les modifications apportées à un ensemble de données grâce aux versionnement,
- **Accessibilité** : permettre aux utilisateurs d'interroger facilement sur les modifications et/ou les corrections apportées.



Pourquoi un service errata pour CMIP6?

Allô Houston ?
J'ai un truc bizarre avec un fichier CMIP6?

Tu as regardé sur errata.es-doc.org ?

Non, c'est quoi ? Le dernier réseau social à la mode ?

Le Service Errata en bref... et avec le soutien de es-doc Earth System Documentation et en collaboration avec le DKRZ

https://errata.es-doc.org/

Stakeholders

- ESGF user reporting
- Data provider - Issue evaluation
- Data(node) manager - Data identification
- ESGF user request

Command-Line Interface

- Authentication via credentials setup
- Local template
- Validation based on JSON schemas

Errata Front-End

- Authentication via Github login
- Web forms
- Browser based validation
- Issue lookup
- Issue viewer
- Version tracker

Errata Web-Service

- Authorization via Github orgs & teams
- Validation based on JSON schemas
- Database storage

Handle Service

- RabbitMQ consumer
- Issue identifier recording
- Persistent IDentifier storage

Le Service Errata en bref... et avec le soutien de es-doc Earth System Documentation et en collaboration avec le DKRZ

https://errata.es-doc.org/

Démo n°1 : Naviguer dans l'errata CMIP6 sans ramer.

https://errata.es-doc.org/

https://es-doc.github.io/esdoc-errata-client/ ➡ Rubrique *"Look up an issue"*

Des données aux multiples "facettes"

Toutes les données d'un même projet sont organisées, nommées et décrites en suivant des règles. L'ensemble de ces règles forme un plan de gestion des données, également appelé "*Data Reference Syntax*" (DRS).

La DRS décrit chaque paramètre permettant de trier les données. Ces paramètres sont appelés **facettes** (ex. institut, modèle, fréquence, etc.).

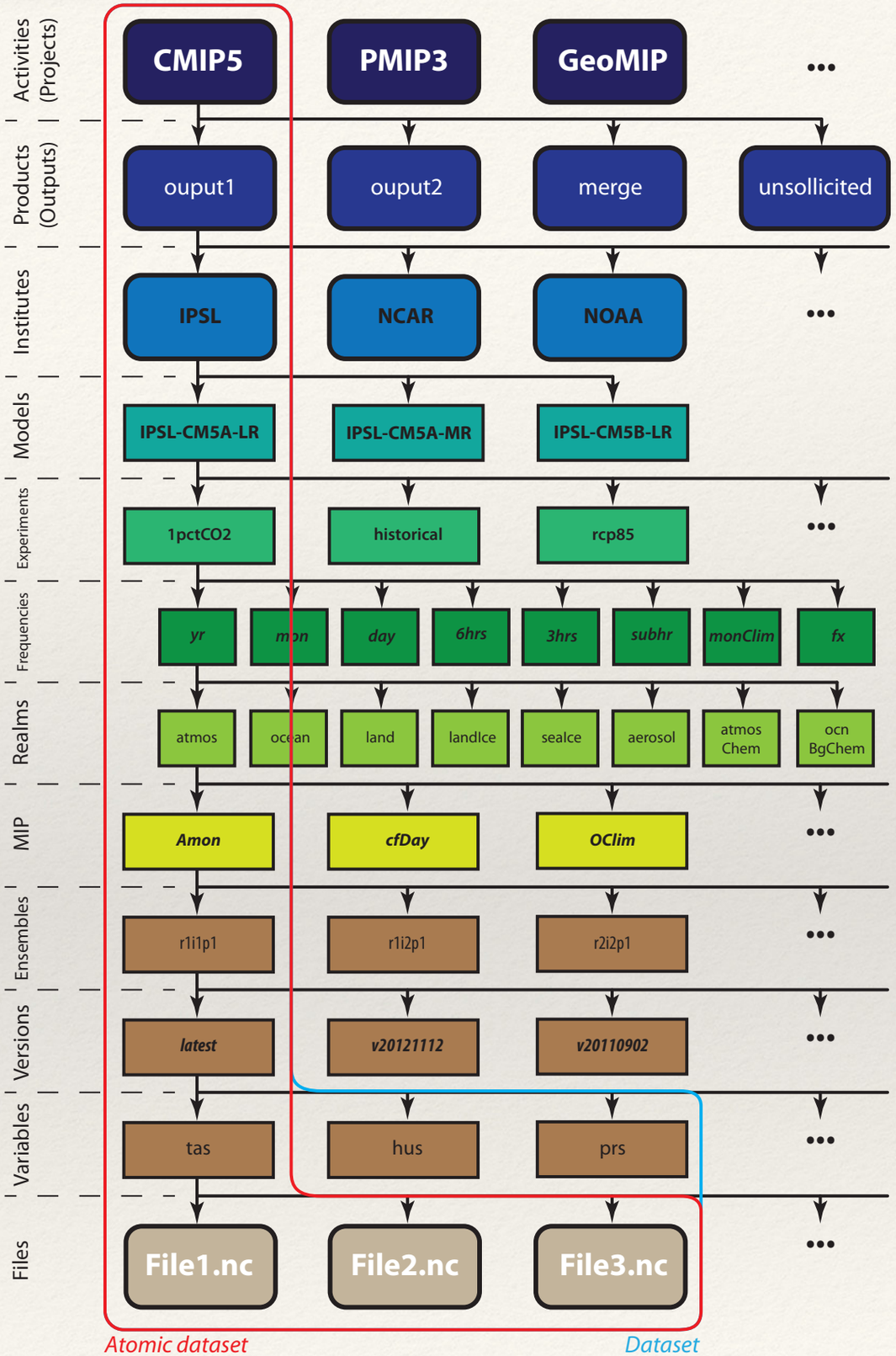
La DRS définit un niveau unitaire dans l'arborescence qui regroupe les fichiers résultant d'une même simulation (i.e., d'un même modèle, d'une même expérience, d'un même membre) ; et qui peut être versionner. Une version de cet ensemble de fichiers est ce que l'on nomme un **dataset**.

Un dataset s'identifie par une succession ordonnée de facettes séparées par des points et qui se termine par un numéro de version.

Exemples:

```
cmip5.output1.IPSL.IPSL-CM5A-LR.
1pctCO2.mon.atmos.Amon.r1i1p1#20110427

CMIP6.CMIP.IPSL.IPSL-CM6A-LR.
1pctCO2.r1i1p1f1.Amon.tas.gr#20180605
```



Vous avez dit *PID* ? Kesako ?

Chaque fichier CMIP6 se voit attribué au moment de sa création une chaîne de caractère immuable et unique. Cet identifiant est enregistré de façon permanente dans une base au moment de la publication du fichier, et ce même si le fichier est ensuite retiré/supprimé. On parle alors de **Persistent IDentifier** (PID) ou encore de ***Handle***. Syntaxiquement un PID est la concaténation d'un identifiant universel unique (ou UUID*) et d'un préfixe définit pour chaque projet, tous deux séparés par un "/".

(*) Un UUID se présente habituellement sous la forme de groupes de caractères hexadécimaux en minuscules séparés par des tirets.

Exemple de PID:

`21.14100/ea6bf619-23fd-4270-9fdc-d89fb3389271`

(`21.14100` est le préfixe défini pour CMIP6)

Le principe du PID se rapproche de celui du DOI (Digital Object Identifier), à la différence qu'un PID n'offre aucune garantie quant à la conservation des données qu'il décrit. Cependant, il restera toujours valide et conduira à au moins une page affichant les informations soumises lors de l'enregistrement de l'identifiant. Cela permet à ESGF de proposer ***des identifiants pour des ensembles de données dont l'état de conservation est encore inconnu***.

Lors de la publication sur ESGF, un PID est également généré pour chaque dataset (c'est-à-dire pour un ensemble de fichier) avec la même syntaxe.

Vous avez dit *PID* ? Kesako ?

```
# PID où te caches-tu ?

$> ncdump -h tas_Amon_IPSL-CM6A-LR_1pctCO2_r1i1p1f1_gr_185001-199912.nc
netcdf tas_Amon_IPSL-CM6A-LR_1pctCO2_r1i1p1f1_gr_185001-199912 {
dimensions:
   axis_nbounds = 2 ;
   lon = 144 ;
   lat = 143 ;
   time = UNLIMITED ; // (1800 currently)
variables:
   float lat(lat) ;
[...]
// global attributes:
     :Conventions = "CF-1.7 CMIP-6.2" ;
     :creation_date = "2018-05-02T17:06:39Z" ;
[...]
     :tracking_id = "hdl:21.14100/ea6bf619-23fd-4270-9fdc-d89fb3389271" ;
[...]


# Astuce

$> ncdump -h tas_Amon_IPSL-CM6A-LR_1pctCO2_r1i1p1f1_gr_185001-199912.nc | grep
tracking_id | cut -d '"' -f 2
hdl:21.14100/ea6bf619-23fd-4270-9fdc-d89fb3389271
```

Démo n°2 : Une analyse sans "fautes"

https://errata.es-doc.org/

https://es-doc.github.io/esdoc-errata-client/ ➡ Rubrique *"Track a version"*

Merci !

- *Développeurs :*
  - *Atef Ben Nasser (CNRS/IPSL - abennasser@ipsl.fr)*
  - *Mark Greenslade (SU/IPSL - momipsl@ipsl.fr)*
  - *Guillaume Levavasseur (SU/IPSL - glipsl@ipsl.fr)*
- *Contributions :*
  - Sébastien *Denvil* (CNRS/IPSL)
  - Merret *Buurman* (DKRZ)
  - Katharina *Berger* (DKRZ)
- *Remerciements :*
  - Martina *Stockhause* (DKRZ)
  - Tobias *Weigel* (DKRZ)
  - Ag *Stephens* (CEDA)
  - Alan *Iwi* (CEDA)

Guillaume Levavasseur

Atef Ben Nasser

Mark Greenslade

https://es-doc.github.io/esdoc-errata-client/