SYNDA
SYNchronizes your DAta

Guillaume Levavasseur[1]
Jérôme Raciazek[1]
Sébastien Denvil[1]

prodiguer

December 10th, 2015

Plan:

- Introduction - Overview
  - History & Position
  - Features
- SYNDA Transfer as an ESGF client
  - Basic use
  - Advanced use
- SYNDA Processing for ESGF replication
  - Trigger post-processing
  - Automated publication
- Conclusion - Workplan
  - CMIP6
  - Automated discovery

Agence Nationale de la Recherche ANR - MODÈLES NUMÉRIQUES - CONVERGENCE

is-enes

cnrs

UPMC SORBONNE UNIVERSITÉS

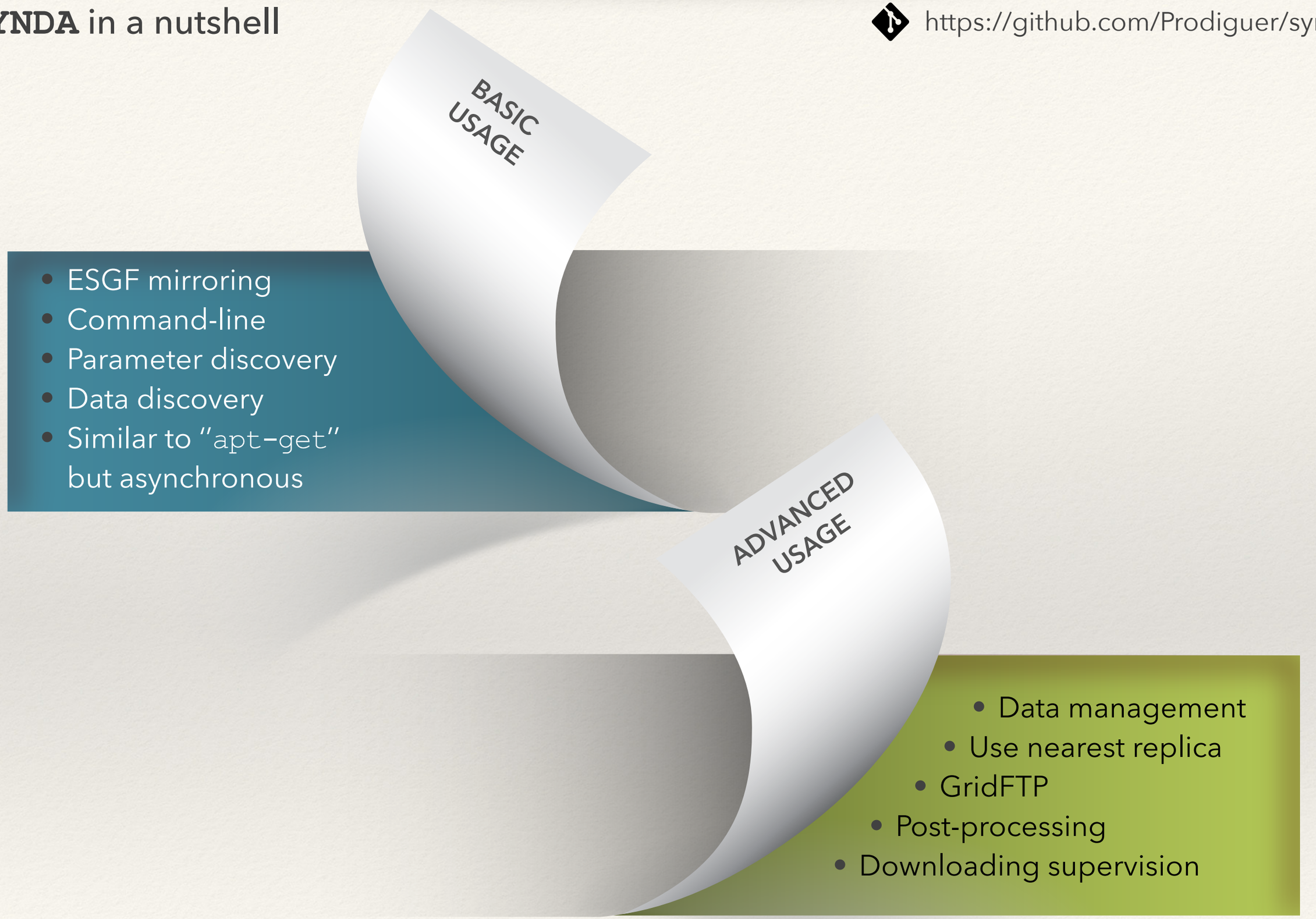
Sciences de l'environnement

Institut Pierre Simon Laplace

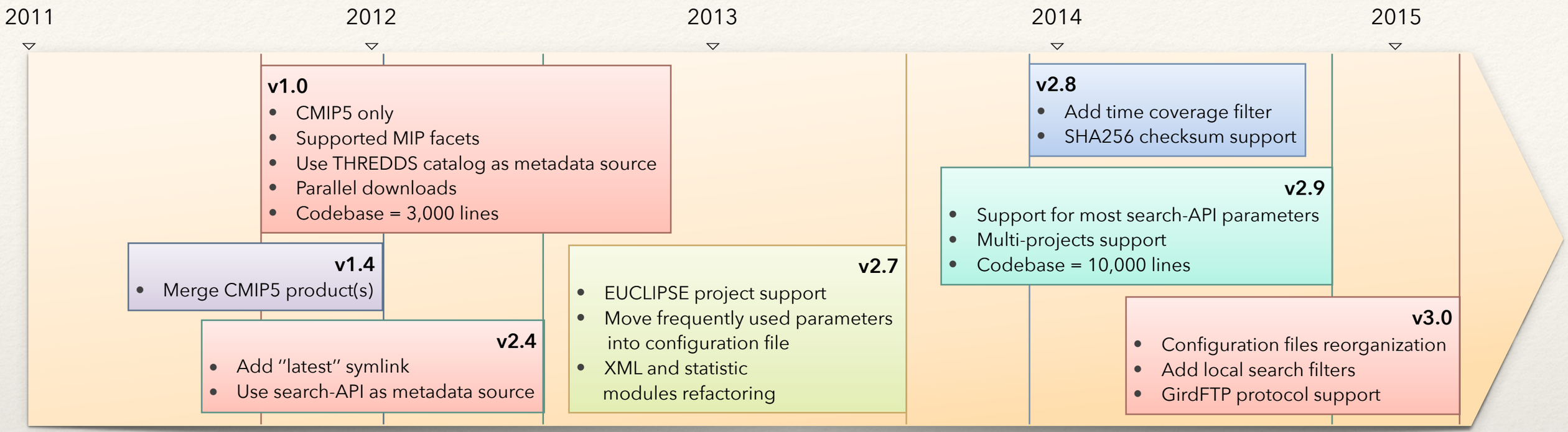
(1) The Institut Pierre Simon Laplace (IPSL) is a research federation including 9 laboratories, 1,400 researchers especially focused on climate and environment.

SYNDA in a nutshell

https://github.com/Prodiguer/synda



SYNDA history



The program will evolve together with the ESGF archive backend functionalities.

SYNDA position in ESGF environment

- ESGF portal (data search)
- `pyclient` (data search)
- `wget` (data download)
- `drs_tool` (versioning and Data Reference Syntax management)

Why SYNDA?

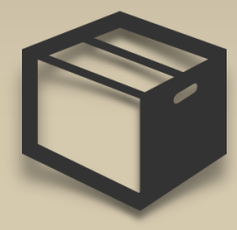
- Helps in building a local mirror, at the intermediary position between ESGF archive and researchers desktop,
- Fast learning curve for user already used to `apt-get` or `yum` tools,
- Facilitate the **distribution**, **access** and **analysis** of international **climate data.**

SYNDA main features

Installation procedure

- DOKER package,
- RPM package,
- `install.sh` script.

System package (v3.1)



- Daemon mode running in background,
- Per-user configuration file (e.g., ~/.syndarc file),
- Online help.

Search

- **Support most ESGF project** (CMIP5, CORDEX, etc.),
- **Facetted data selection** (experiment, variable, etc.),
- **Incremental search** (download only what's new),
- **Nearest replica selection**.

Download

- **Parallel downloads** (using HTTP or GridFTP protocol),
- Checksums control,
- Transfer priority and statistics,
- **Transparent x509 certificate renewal.**

Data management

- **Default storage following DRS tree format,**
- **Template based data management**,
- **Auto-remove old dataset versions**.

SYNDA basic usage: How to download ESGF data in six steps?

```
# Installation
> wget http://dods.ipsl.jussieu.fr/jripsl/synda/synda-3.1.x86_64.rpm
> sudo yum localinstall synda-3.1.x86_64.rpm

# Configuration
> sudo vi /etc/synda/sdt/sdt.conf                          # Set openID and password

# Explore facets using param command
> synda param                                              # List all facets name
> synda param realm                                        # List all values of one facet
> synda param model MPI                                    # Focused list for one facet

# Discover ESGF data using search command
> synda search CMIP5 -d                                    # Search for datasets (default)
> synda search mon tasmax -f                               # Search for files
> synda search CMIP5 va day -v                             # Search for variables
> synda search CORDEX historical r0i0p0 -a                 # Search for aggregations
> synda search day pr -r                                   # Search for replicas

# Request installation
> synda install <FACETS> […]

# Data default location data
> cd /srv/synda/sdt/<PROJECT_DRS>
```

SYNDA advanced usage: Download queue interaction & data management

```
# Display HTTP requests
> synda search <FACET> […] --dry_run

# Print current downloads
> synda watch

# Print download queue
> synda queue

# Retry transfers in error
> synda retry
```

```
# Show dataset details (<DATASET> = <DATASET_ID>.<VERSION>)
> synda show <DATASET>

# List all dataset versions
> synda version <DATASET>

# List local files
> synda list -f

# Update datasets to new version
> synda upgrade

# Remove a dataset from SYNDA database and your local filesystem
> synda remove <DATASET>

# Remove old dataset versions
> synda autoremove

# Show history (add/remove)
> synda history
```

SYNDA: The use of template

SYNDA allows to define *templates*. A *template* is a text file that stores data requests details (variables, frequencies, experiments, etc.). An inheritance system allows to share facet values across different templates (e.g., default values can be set for all projects or separately for each project). To explore the ESGF archive with the template, use SYNDA "`-s`" option. SYNDA may be run regularly to download all new files regarding to the defined templates.

```
# Declare a template in sdt/selection folder
> vi /etc/synda/sdt/selection/my_template.txt
```

```
# My SYNDA template

project=CMIP5
experiments=historical amip
models=IPSL-CM5A-LR CNRM-CM5
ensembles=all
variables[atmos][mon]=tas
variables[atmos][3hr]=cltc zfull
variables[land][fx]=sftgif
variables[seaIce][mon]=sic evap
searchapi_host=esgf-node.ipsl.fr
protocol=http
```

```
# Engage the template for discovery and download
> synda install -s my_template.txt
```

SYNDA: Trigger a post-processing

REQUEST

Search criteria called facets are used to select which files to download. They can be set on command line or using a template.

ESGF NODES

SDT retrives the certificates and builds the HTTP requests to Solr corresponding to the search criteria.

SYNDA Transfer

A command line tool to discover and download files from the Earth System Grid Federation (ESGF) archive.

SYNDA Processing

A pipeline for a parallel, orderly and asynchronous post-processing on downloaded NetCDF files.

WORKERS

A Python daemon deals with the database using a Remote Procedure Call (RPC) client. A "worker" is able to run Python or shell scripts.

FILESYSTEM

ESGF files are downloaded using the HTTP or GridFTP protocol and managed on the local filesystem following the Data Reference Syntax.

SDT DATABASE

A SQLite database records each downloaded file and dataset. A complete dataset triggers a "dataset_complete" event, which informs the SDP module to start the pipeline.

SDP DATABASE

A SQLite database describes and follows the post-processing progress of an atomic dataset or a dataset.

SYNDA: How to automate publication for replication?

CRONTAB

ESGF discovery is regularly run to download possible new files according to the installed templates.

ESGF NODES

SDT retrives the certificates and builds the HTTP requests to Solr corresponding to the search criteria.

SYNDA Transfer

A command line tool to discover and download files from the Earth System Grid Federation (ESGF) archive.

FILESYSTEM

ESGF files are downloaded using the HTTP or GridFTP protocol and managed on the local filesystem following the Data Reference Syntax.

SDT DATABASE

A SQLite database records each downloaded file and dataset. A complete dataset triggers a "dataset_complete" event, which informs the SDP module to start the pipeline.

SYNDA Processing

A "publication pipeline" orderly runs the mapfile generation of downloaded new dataset versions, then publishes them as replicas.

ESGF PUBLISHER

A SDP "worker" runs a shell script that embeds the `esgpublish` command lines with the appropriate flags (e.g., `--set-replicas`).

MAPFILE

A SDP "worker" runs a shell script that embeds the `esgscan_directory` command line. One mapfile per dataset is automatically generated.

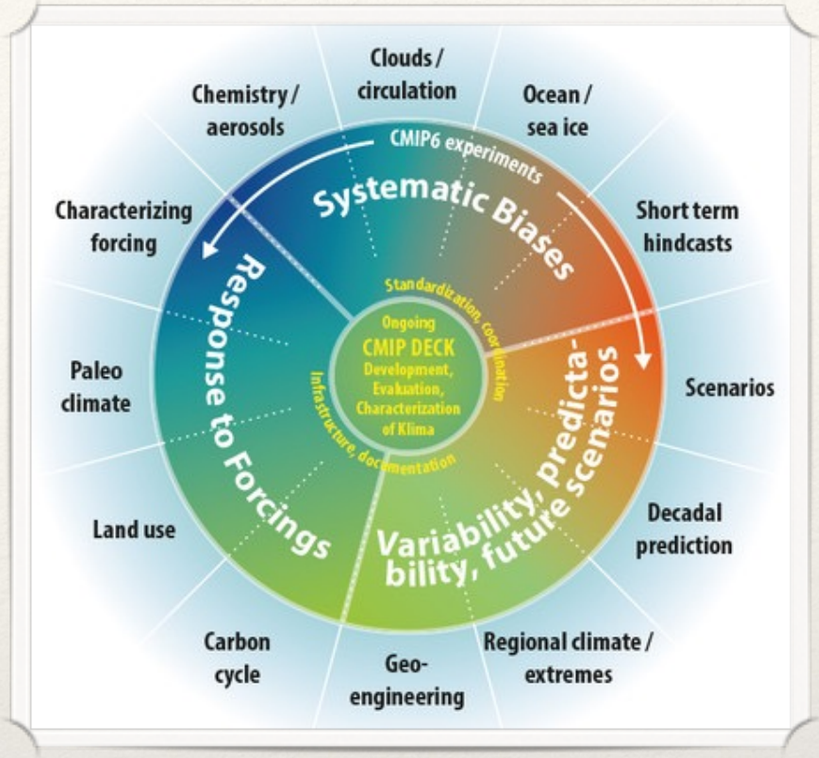
SDP DATABASE

A SQLite database describes and follows the post-processing progress of an atomic dataset or a dataset.

SYNDA workplan

CMIP6 HORIZON: INCREASING DATA AMOUNT AND DATA MANAGEMENT

No fast replica synchronization or notification mechanisms are currently supported. Thereby in general original data can be unpublished or modified without effects on replica sites. **Automatic replication procedures are considered essential for CMIP6 data.**



See ESGF WIP paper on "Versioning and Replication for CMIP6"

OUR GOAL: REDUCE THE SYNCHRONISATION TIME BETWEEN ESGF ARCHIVE AND THE LOCAL FILESYSTEM

Transition from an offline "on demand" discovery, to an online automatic discovery is crucial. We aims to work on the ESGF "hooks" necessary to support fully automatic replication procedures in the future. Several ways are investigated to automatically trigger the SYNDA discovery:

- Scheduled high frequency pull (i.e., crontab),
- Push to SYNDA (but requires security layer),
- Subscribe to an ESGF feed (e.g., RSS),
- Other?

Thank you for your attention.

- *Developer:*
  - Jérôme Raciazek (jripsl@ipsl.jussieu.fr)
- *Contributors:*
  - Sébastien Denvil (sdipsl@ipsl.jussieu.fr)
  - Guillaume Levavasseur (glipsl@ipsl.juisseu.fr)

https://github.com/Prodiguer/synda