www.bsc.es

Lisboa, 27/09/2016

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

Online metadata generation with CMOR

Joint IS-ENES Workshop on Workflows and Metadata Generation

Pierre-Antoine Bretonnière

Plan

- Why metadata?
- Introduction to the BSC workflow
  - BSC ecosystem
  - Autosubmit
- CMOR
  - Concept
  - CMIP format
- Online generation and integration of metadata
  - Where to integrate the metadata?
  - From where?
- Future plans

Why metadata?

- Keep track of your own work
- Help sharing your data
- **Improve reproducibility**

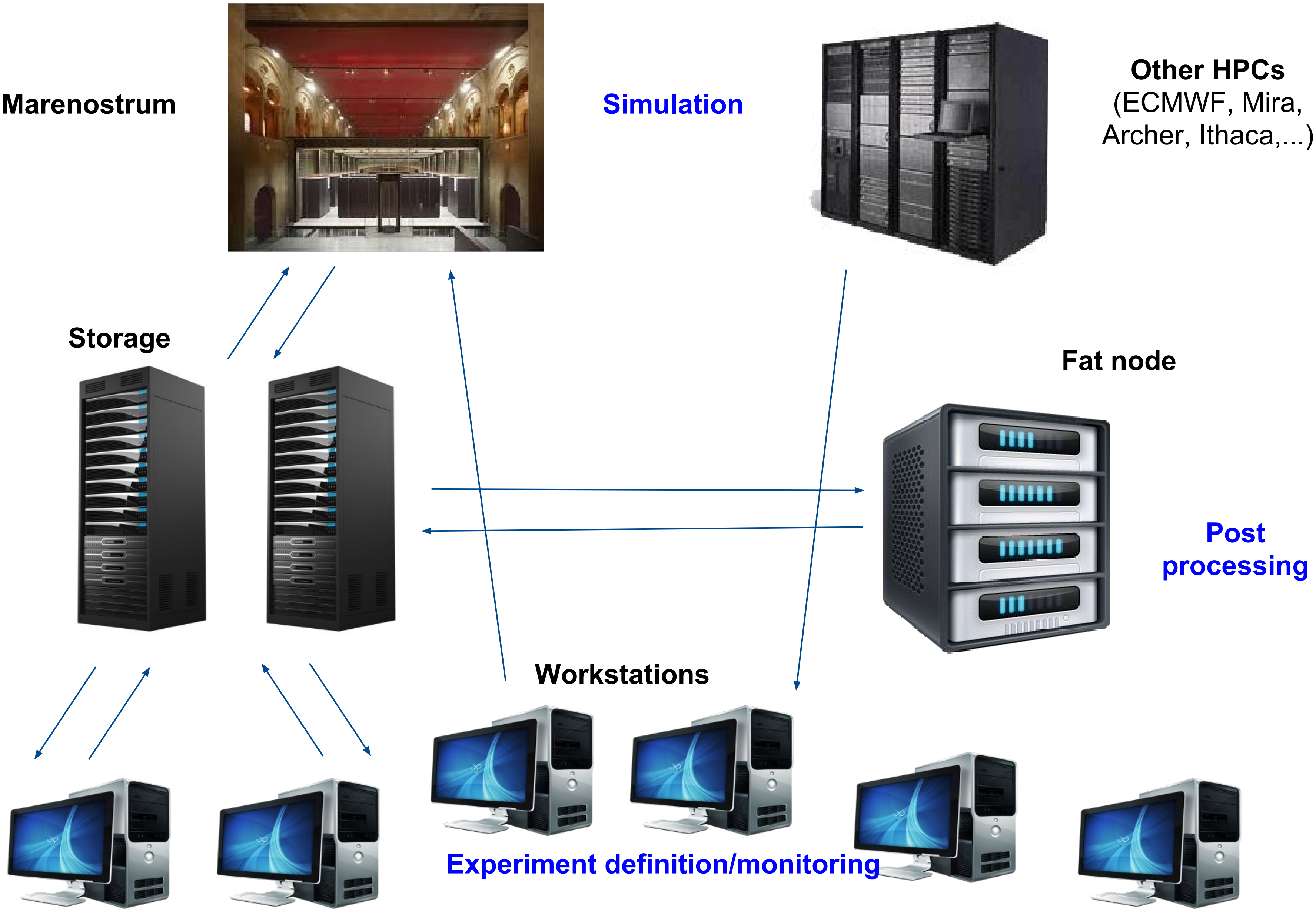
What metadata?

How to improve the reproducibility?

- Generating metadata for as many things as possible:
  - Extensive description of the experiment setup (model, initialization, physics, forcings, start dates,...)
  - Precise physical description of the variables (long_names, units, cell_methods,...)
  - Software versions (git tags, branch, commit id,...)
  - Software dependencies (git submodules)
  - Creation dates
  - Unique identifiers…
  - Physical contact points (people)

- Making it as “user-friendly” and automatic as possible

The BSC ecosystem



Marenostrum

Simulation

Other HPCs
(ECMWF, Mira, Archer, Ithaca,...)

Storage

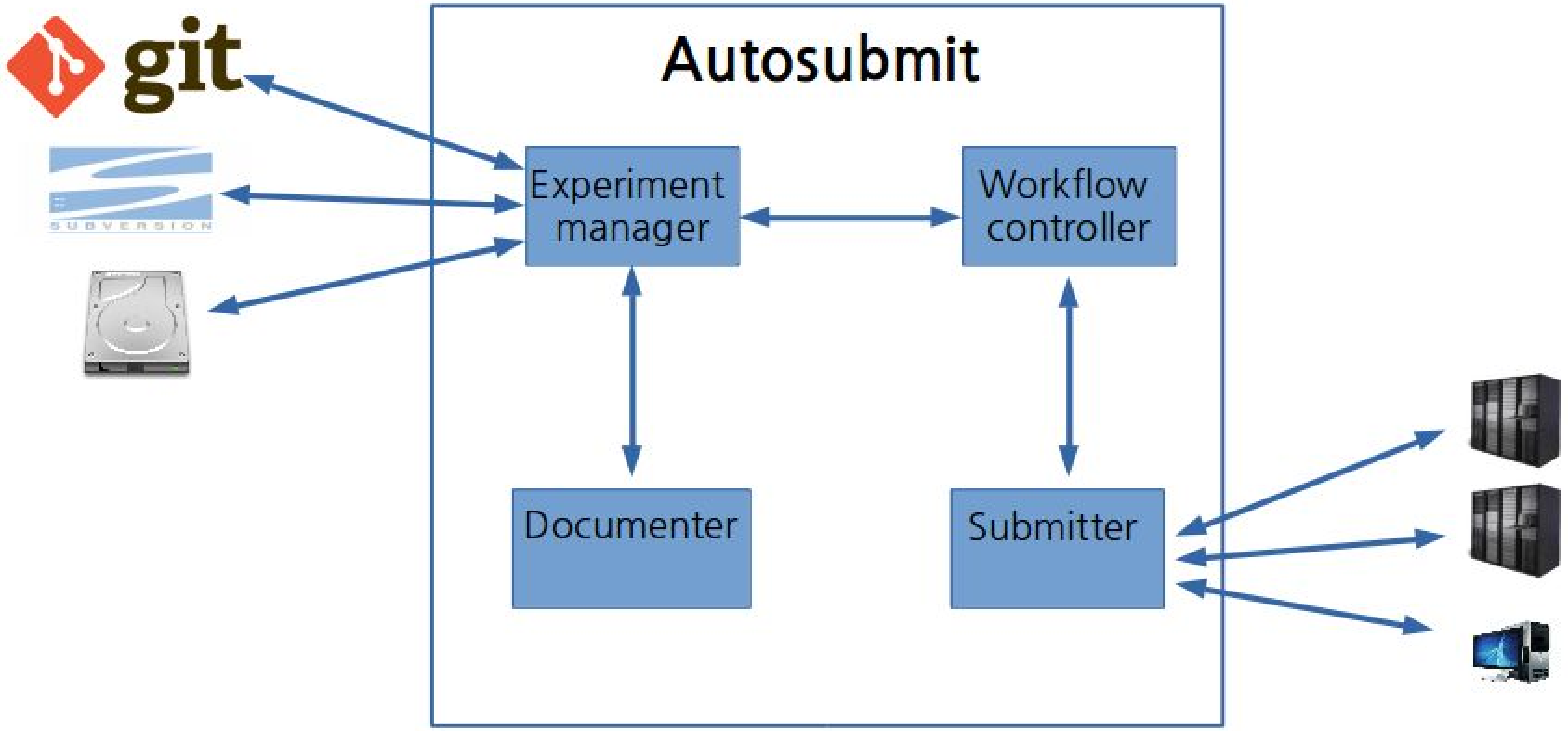
Fat node

Post processing

Workstations

Experiment definition/monitoring

Autosubmit



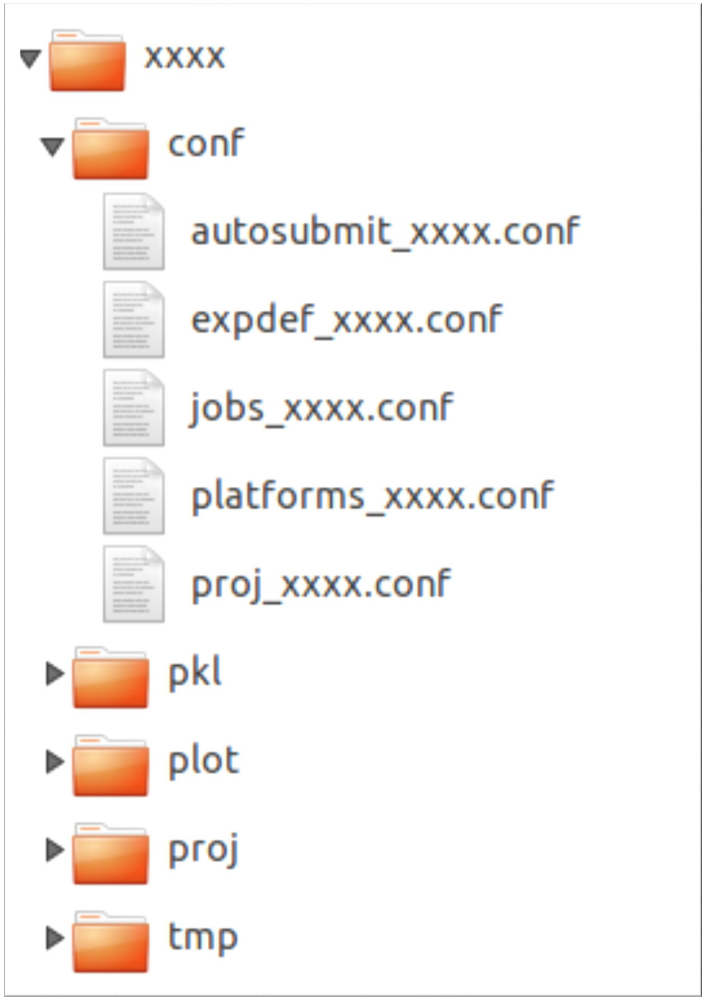
https://pypi.python.org/pypi/autosubmit

Tutorial on Autosubmit tomorrow at 11.30

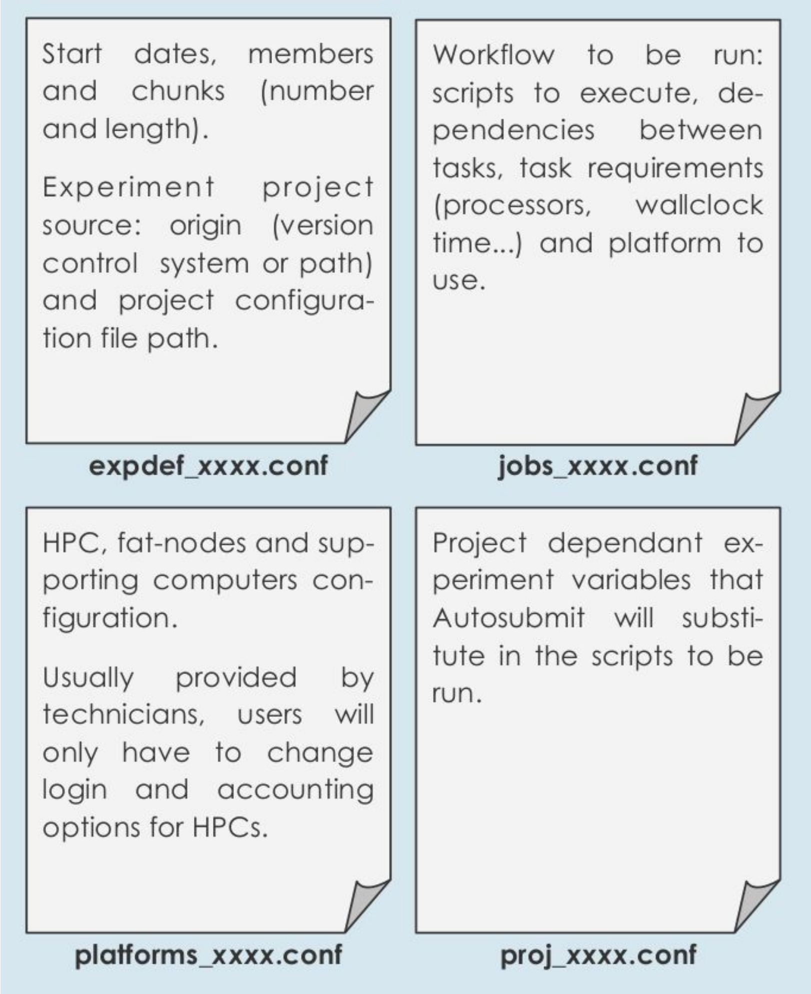
Autosubmit

```
autosubmit expid –H HPCname
```

```
autosubmit create xxxx
```



- xxxx
  - conf
    - autosubmit_xxxx.conf
    - expdef_xxxx.conf
    - jobs_xxxx.conf
    - platforms_xxxx.conf
    - proj_xxxx.conf
  - pkl
  - plot
  - proj
  - tmp



expdef_xxxx.conf

Start dates, members and chunks (number and length).

Experiment project source: origin (version control system or path) and project configuration file path.

jobs_xxxx.conf

Workflow to be run: scripts to execute, dependencies between tasks, task requirements (processors, wallclock time...) and platform to use.

platforms_xxxx.conf

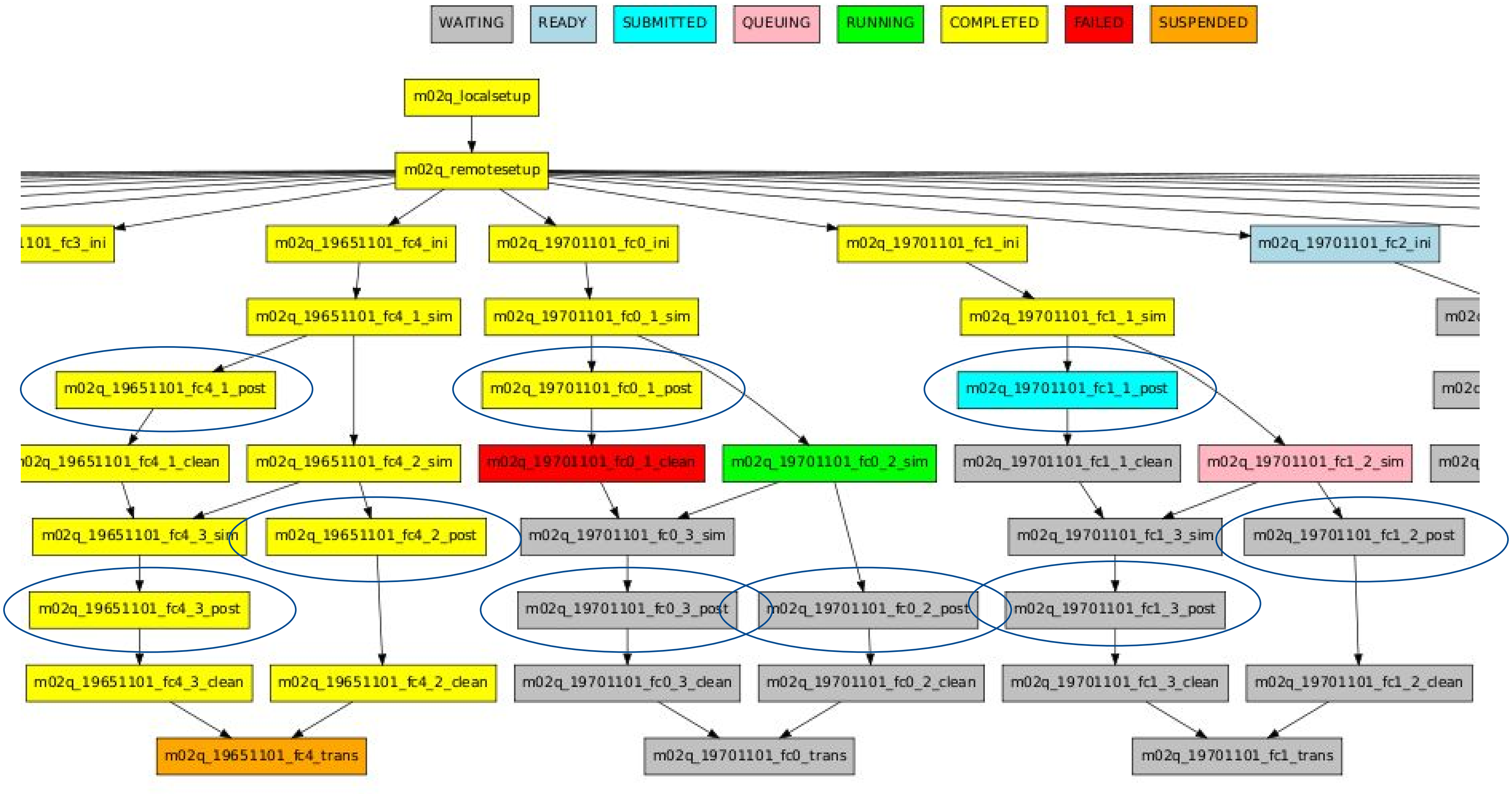
HPC, fat-nodes and supporting computers configuration.

Usually provided by technicians, users will only have to change login and accounting options for HPCs.

proj_xxxx.conf

Project dependant experiment variables that Autosubmit will substitute in the scripts to be run.

Online CMORization



CMOR: what is it?

- **C**limate **M**odel **O**utput **R**ewriter

- C library developed by Lawrence Livermore National Library (LLN)

- Version 2 used for CMIP5/SPECS/CORDEX

- Version 3 just released for CMIP6 with
  - New Data Reference Syntax (DRS)
  - Json Model Intercomparison Project (MIPs)
  - More modularity

CMOR metadata: example of CMIP6

- **Directory name:**

  <mip_era>/<institute_id>/<source_id>/<activity_id>/<experiment_id>/<variant_label>/<table>/<variable_id>/<grid_label>/<version>

  CMIP6/BSC-CNS/EC-Earth/DCPP/histSST/r1i1p1/CMIP6_day/tas/gn/v1/

- **File name:**

  <variable_id>_<table>_<experiment_id>_<source_id>_<variant_label>_<grid_label>_<date>.nc

  tas_CMIP6_day_histSST_EC-Earth_r1i1p1_gn_1980101-19810131.nc

CMOR metadata: example of CMIP6

Global attributes

- **CMOR mandatory:**

Variant_label, activity_id, branch_method, Conventions, creation_date, mip_era, data_specs_version, experiment_id, experiment, forcing_index, further_info_url, frequency, grid, grid_label, grid_resolution, initialization_index, institution, institution_id, license, physics_index, product, realization_index, realm, variant_label, source, source_id, source_type, sub_experiment, sub_experiment_id, table_id, tracking_id, variable_id

- **BSC adds-on:**

Autosubmit version and model, modules tags

CMOR: what is it?

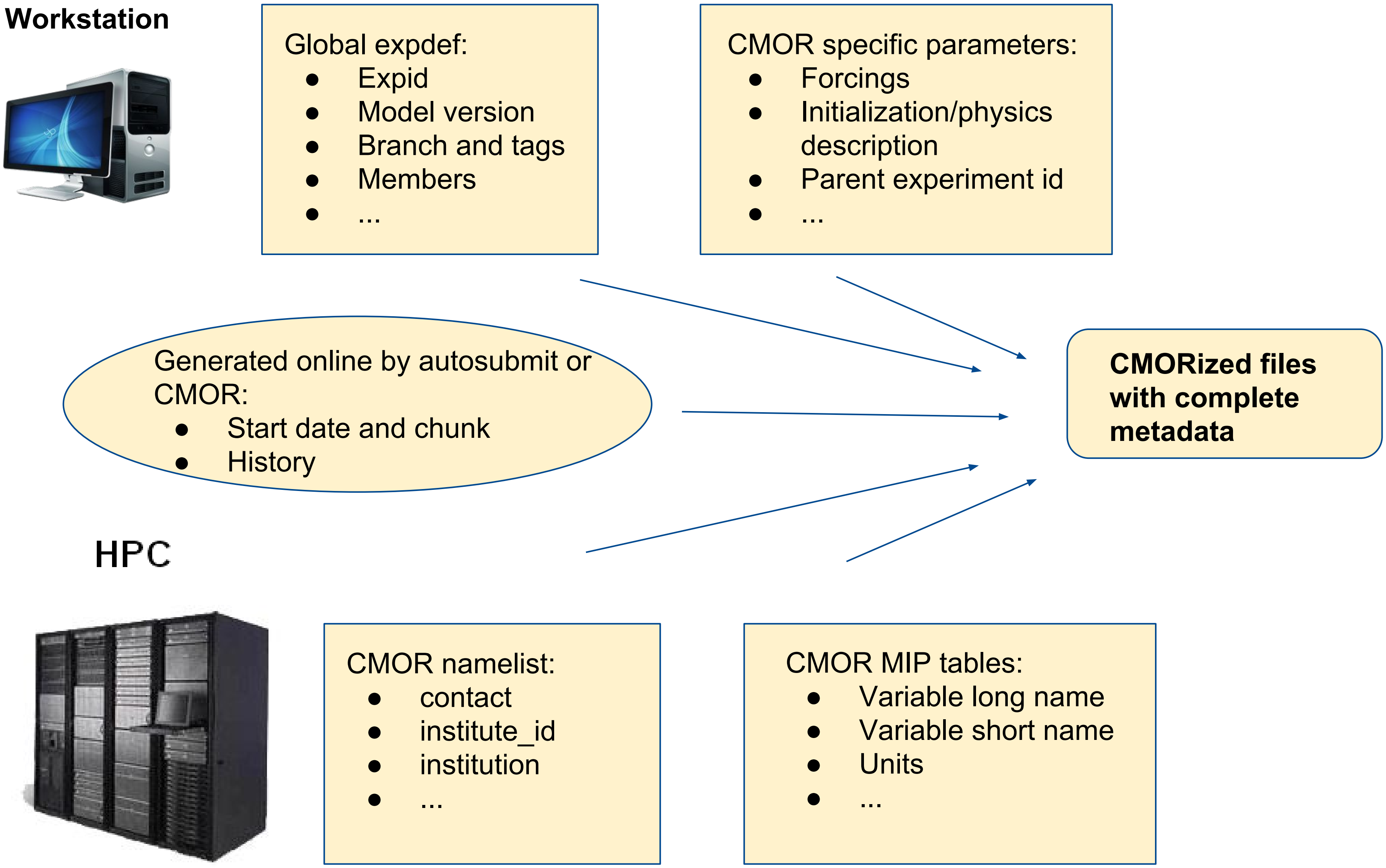
- (Re)writes raw outputs of the models with names that comply with the project conventions (CMIP5, CORDEX, SPECS, CMIP6)
- **Fills in the metadata required by the project**
- Programs work with **external namelist** containing the metadata

CMOR: what is it?

- (Re)writes raw outputs of the models with names that comply with the project conventions (CMIP5, CORDEX, SPECS, CMIP6)
- **Fills in the metadata required by the project**
- Programs work with **external namelist** containing the metadata

=> how to fill in this namelist as automatically as possible?

Metadata provenance



Workstation

Global expdef:
- Expid
- Model version
- Branch and tags
- Members
- ...

CMOR specific parameters:
- Forcings
- Initialization/physics description
- Parent experiment id
- ...

Generated online by autosubmit or CMOR:
- Start date and chunk
- History

CMORized files with complete metadata

HPC

CMOR namelist:
- contact
- institute_id
- institution
- ...

CMOR MIP tables:
- Variable long name
- Variable short name
- Units
- ...

“Offline” CMORization

- Online CMORization added in the workflow when experiments had already run: what do we do with these?

↓

- Earth diagnostics automatically run a CMOR-like script for both ocean and atmospheric variables.

- Metadata picked in autosubmit configuration files or automatically asked to the user a posteriori

✓

Global expdef:

- Expid
- Model version
- Branch and tags
- Members
- ...

CMOR specific parameters:

- Forcings
- Initialization/physics description
- Parent experiment id
- ...

✗

Ongoing and future plans

- Add ES-DOC in Autosubmit
- Modularize CMORization process
- Add complete history of file processing all along its life to keep track of the changes
- Use of a community (EC-Earth) level common CMORization tools

Q & A

www.bsc.es

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

Thank you!

For further information please contact
pierre-antoine.bretonniere@bsc.es