BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

BSC data organization

BSC-Ouranos talk

PA Bretonnière

27/10/2021

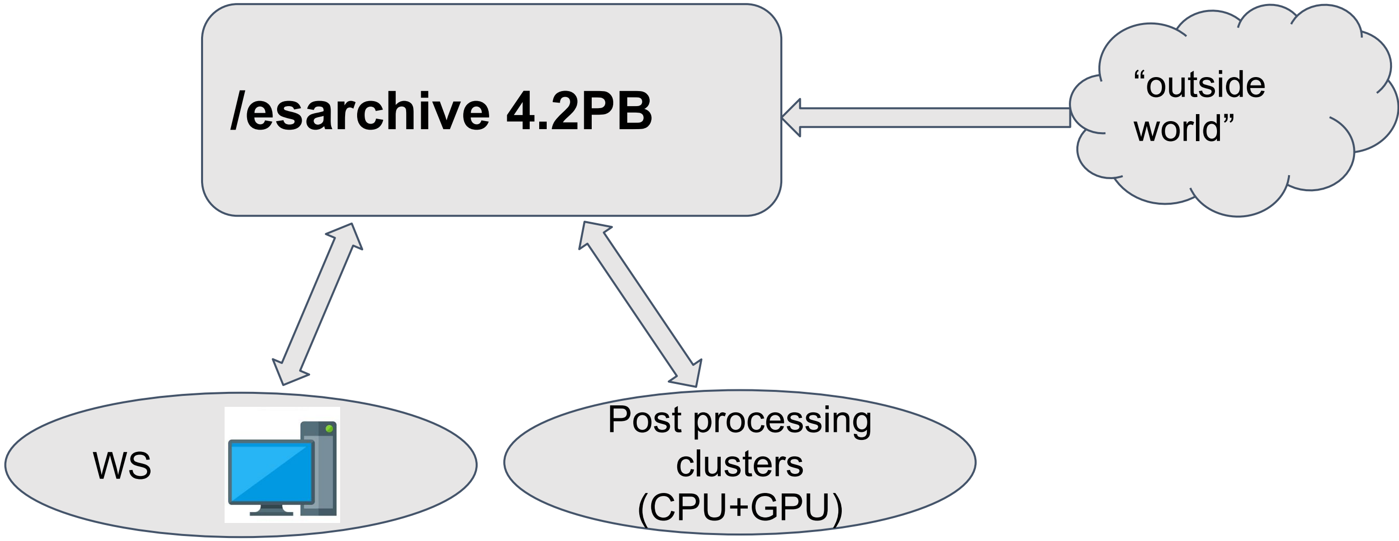
Outline

- General overview
- Conventions and organization
- Requests
- Data formatting tools and checkers
- Link with the (department) tools
- Conclusions and main challenges of data organization

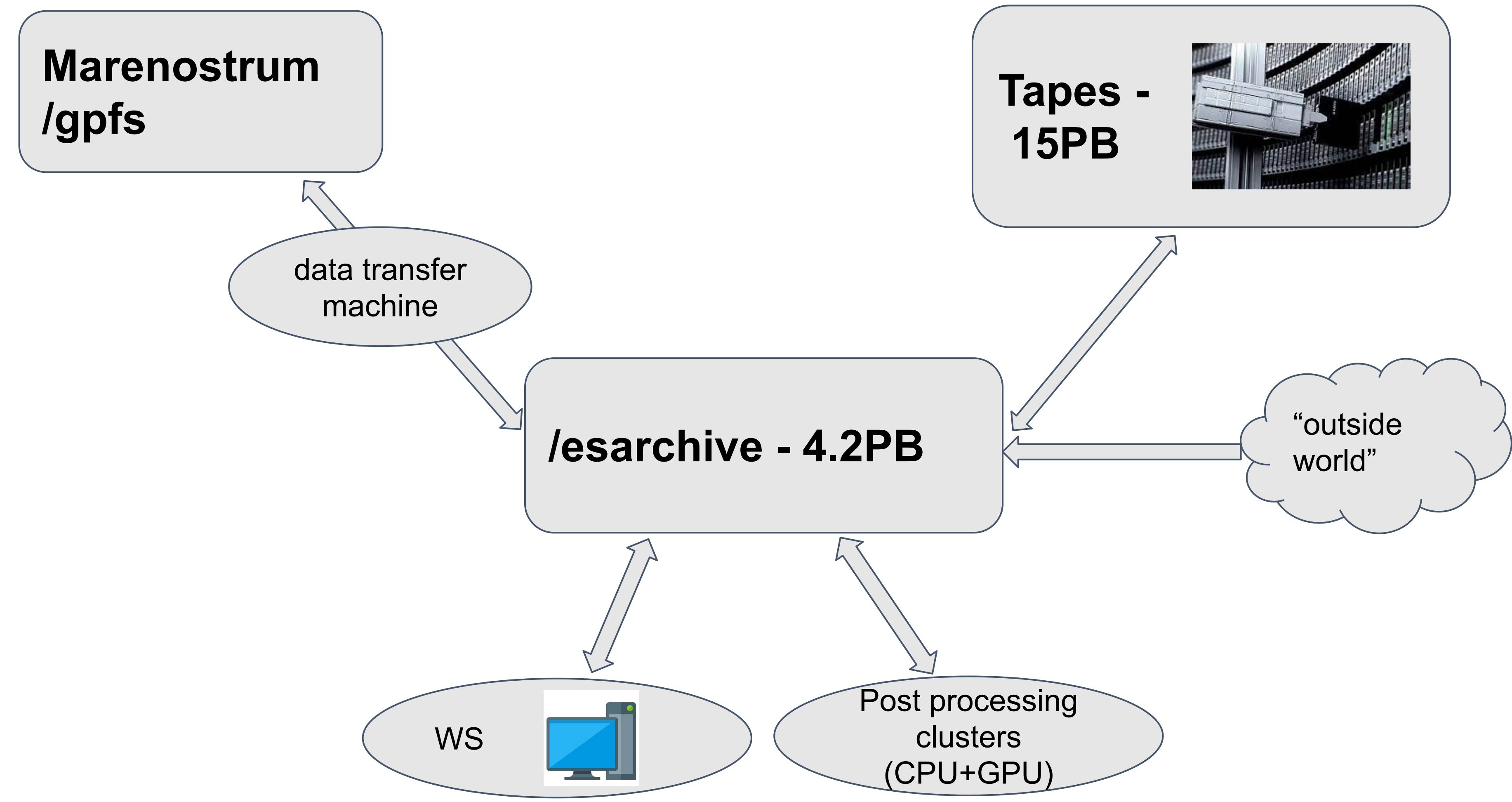
General overview

- **BSC:** 4 scientific departments: Life, CASE, Computer, **Earth**
- **Earth sciences department:**
    - ~105 people, from which ~80 use or generate data daily
    - 4 groups: Climate Prediction, Atmospheric Composition, Earth System Services, Computational Earth Sciences
    - people have different needs, scientific/technical background
    - common storage with 4PB of data for everybody
- **Objective:** Curate, optimize and develop this common data pool for best use for internal usage
- 1 full time person + 3 part time to manage it

General overview



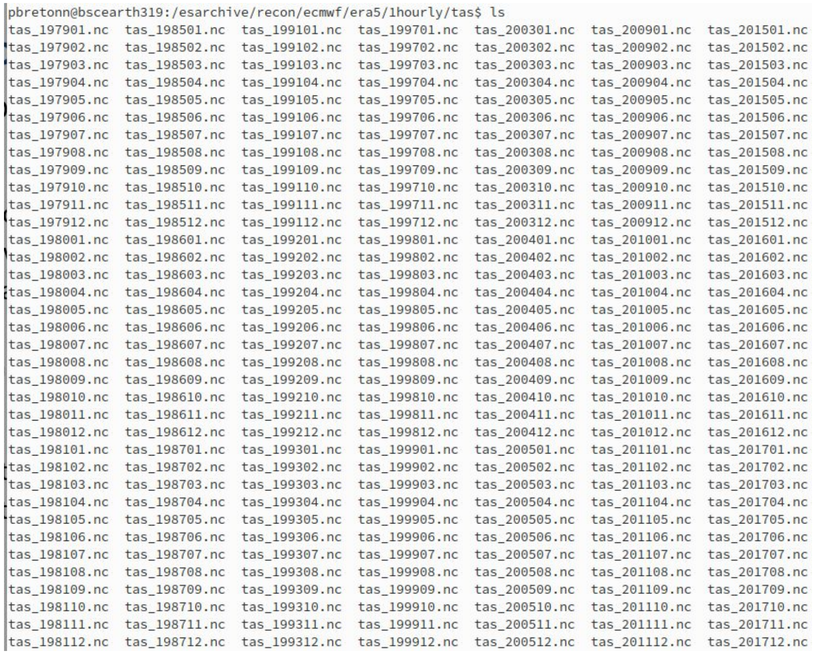
General overview



Conventions and organization

- Store the data in netCDF (+grib in some cases)

- /esarchive/[exp-recon-obs]/$institute/$dataset/ *(/esarchive/recon/ecmwf/era5)*
  - $frequency/$variable [-$grid] *(hourly/tas)*
  - original_files
  - scripts (version controlled under Gitlab)



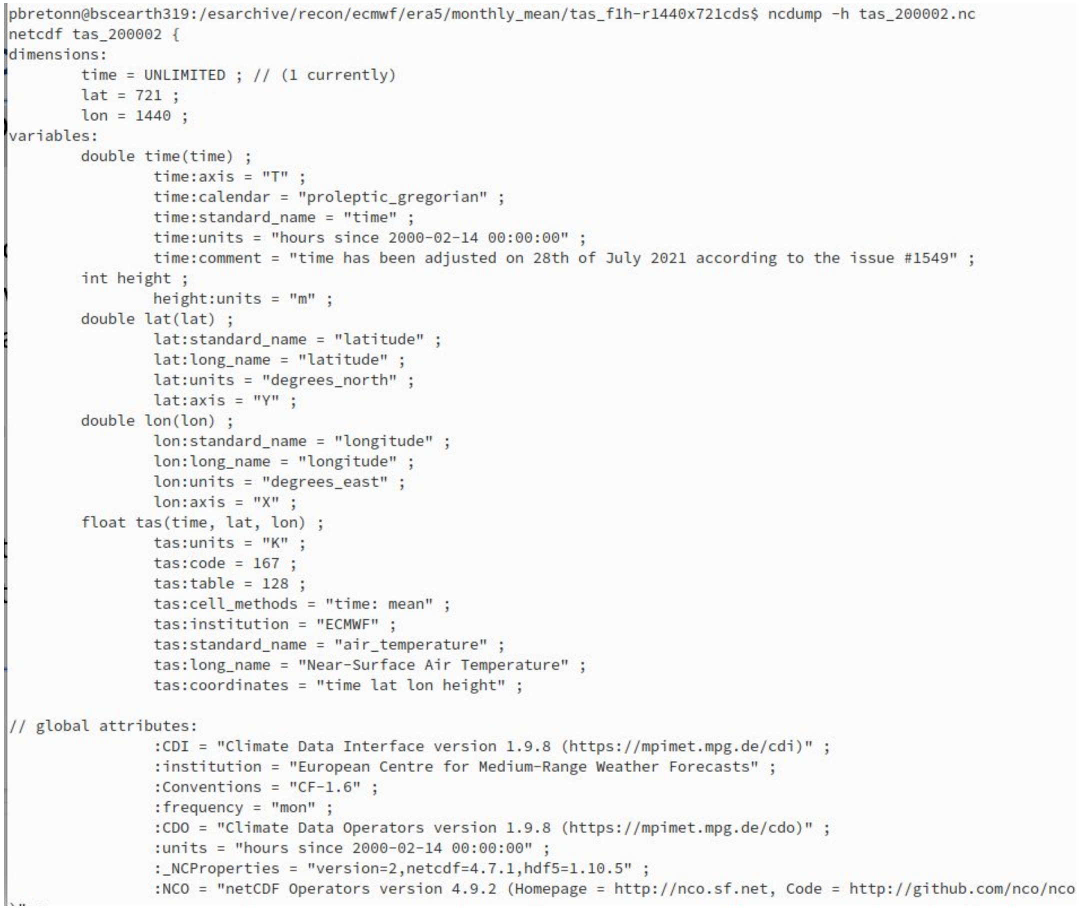
```
pbretonn@bscearth319:/esarchive/recon/ecmwf/era5/1hourly/tas$ ls
tas_197901.nc  tas_198501.nc  tas_199101.nc  tas_199701.nc  tas_200301.nc  tas_200901.nc  tas_201501.nc
tas_197902.nc  tas_198502.nc  tas_199102.nc  tas_199702.nc  tas_200302.nc  tas_200902.nc  tas_201502.nc
tas_197903.nc  tas_198503.nc  tas_199103.nc  tas_199703.nc  tas_200303.nc  tas_200903.nc  tas_201503.nc
tas_197904.nc  tas_198504.nc  tas_199104.nc  tas_199704.nc  tas_200304.nc  tas_200904.nc  tas_201504.nc
tas_197905.nc  tas_198505.nc  tas_199105.nc  tas_199705.nc  tas_200305.nc  tas_200905.nc  tas_201505.nc
tas_197906.nc  tas_198506.nc  tas_199106.nc  tas_199706.nc  tas_200306.nc  tas_200906.nc  tas_201506.nc
tas_197907.nc  tas_198507.nc  tas_199107.nc  tas_199707.nc  tas_200307.nc  tas_200907.nc  tas_201507.nc
tas_197908.nc  tas_198508.nc  tas_199108.nc  tas_199708.nc  tas_200308.nc  tas_200908.nc  tas_201508.nc
tas_197909.nc  tas_198509.nc  tas_199109.nc  tas_199709.nc  tas_200309.nc  tas_200909.nc  tas_201509.nc
tas_197910.nc  tas_198510.nc  tas_199110.nc  tas_199710.nc  tas_200310.nc  tas_200910.nc  tas_201510.nc
tas_197911.nc  tas_198511.nc  tas_199111.nc  tas_199711.nc  tas_200311.nc  tas_200911.nc  tas_201511.nc
tas_197912.nc  tas_198512.nc  tas_199112.nc  tas_199712.nc  tas_200312.nc  tas_200912.nc  tas_201512.nc
tas_198001.nc  tas_198601.nc  tas_199201.nc  tas_199801.nc  tas_200401.nc  tas_201001.nc  tas_201601.nc
tas_198002.nc  tas_198602.nc  tas_199202.nc  tas_199802.nc  tas_200402.nc  tas_201002.nc  tas_201602.nc
tas_198003.nc  tas_198603.nc  tas_199203.nc  tas_199803.nc  tas_200403.nc  tas_201003.nc  tas_201603.nc
tas_198004.nc  tas_198604.nc  tas_199204.nc  tas_199804.nc  tas_200404.nc  tas_201004.nc  tas_201604.nc
tas_198005.nc  tas_198605.nc  tas_199205.nc  tas_199805.nc  tas_200405.nc  tas_201005.nc  tas_201605.nc
tas_198006.nc  tas_198606.nc  tas_199206.nc  tas_199806.nc  tas_200406.nc  tas_201006.nc  tas_201606.nc
tas_198007.nc  tas_198607.nc  tas_199207.nc  tas_199807.nc  tas_200407.nc  tas_201007.nc  tas_201607.nc
tas_198008.nc  tas_198608.nc  tas_199208.nc  tas_199808.nc  tas_200408.nc  tas_201008.nc  tas_201608.nc
tas_198009.nc  tas_198609.nc  tas_199209.nc  tas_199809.nc  tas_200409.nc  tas_201009.nc  tas_201609.nc
tas_198010.nc  tas_198610.nc  tas_199210.nc  tas_199810.nc  tas_200410.nc  tas_201010.nc  tas_201610.nc
tas_198011.nc  tas_198611.nc  tas_199211.nc  tas_199811.nc  tas_200411.nc  tas_201011.nc  tas_201611.nc
tas_198012.nc  tas_198612.nc  tas_199212.nc  tas_199812.nc  tas_200412.nc  tas_201012.nc  tas_201612.nc
tas_198101.nc  tas_198701.nc  tas_199301.nc  tas_199901.nc  tas_200501.nc  tas_201101.nc  tas_201701.nc
tas_198102.nc  tas_198702.nc  tas_199302.nc  tas_199902.nc  tas_200502.nc  tas_201102.nc  tas_201702.nc
tas_198103.nc  tas_198703.nc  tas_199303.nc  tas_199903.nc  tas_200503.nc  tas_201103.nc  tas_201703.nc
tas_198104.nc  tas_198704.nc  tas_199304.nc  tas_199904.nc  tas_200504.nc  tas_201104.nc  tas_201704.nc
tas_198105.nc  tas_198705.nc  tas_199305.nc  tas_199905.nc  tas_200505.nc  tas_201105.nc  tas_201705.nc
tas_198106.nc  tas_198706.nc  tas_199306.nc  tas_199906.nc  tas_200506.nc  tas_201106.nc  tas_201706.nc
tas_198107.nc  tas_198707.nc  tas_199307.nc  tas_199907.nc  tas_200507.nc  tas_201107.nc  tas_201707.nc
tas_198108.nc  tas_198708.nc  tas_199308.nc  tas_199908.nc  tas_200508.nc  tas_201108.nc  tas_201708.nc
tas_198109.nc  tas_198709.nc  tas_199309.nc  tas_199909.nc  tas_200509.nc  tas_201109.nc  tas_201709.nc
tas_198110.nc  tas_198710.nc  tas_199310.nc  tas_199910.nc  tas_200510.nc  tas_201110.nc  tas_201710.nc
tas_198111.nc  tas_198711.nc  tas_199311.nc  tas_199911.nc  tas_200511.nc  tas_201111.nc  tas_201711.nc
tas_198112.nc  tas_198712.nc  tas_199312.nc  tas_199912.nc  tas_200512.nc  tas_201112.nc  tas_201712.nc
```

- 1 variable, 1 month (/start date) per file
- + some links for projects (CMIP)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Conventions and organization

- short names based on CMOR/CMIP6 conventions (tos,tas,tasmin,...)
- CF compliant (standard and long names, units, dimensions,...)
- homogeneous time axis
- coordinates

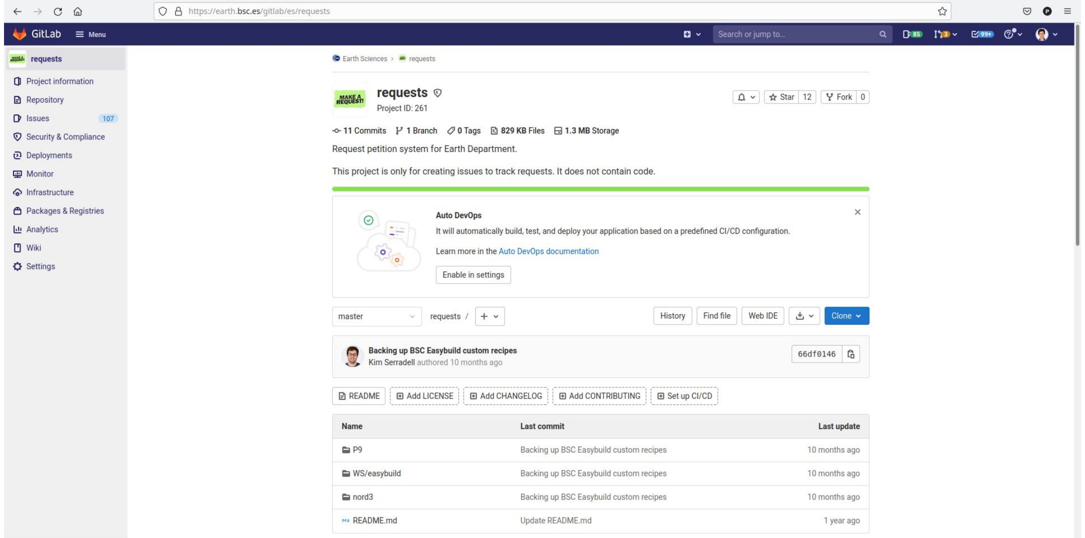


```
pbretonn@bscearth319:/esarchive/recon/ecmwf/era5/monthly_mean/tas_f1h-r1440x721cds$ ncdump -h tas_200002.nc
netcdf tas_200002 {
dimensions:
        time = UNLIMITED ; // (1 currently)
        lat = 721 ;
        lon = 1440 ;
variables:
        double time(time) ;
                time:axis = "T" ;
                time:calendar = "proleptic_gregorian" ;
                time:standard_name = "time" ;
                time:units = "hours since 2000-02-14 00:00:00" ;
                time:comment = "time has been adjusted on 28th of July 2021 according to the issue #1549" ;
        int height ;
                height:units = "m" ;
        double lat(lat) ;
                lat:standard_name = "latitude" ;
                lat:long_name = "latitude" ;
                lat:units = "degrees_north" ;
                lat:axis = "Y" ;
        double lon(lon) ;
                lon:standard_name = "longitude" ;
                lon:long_name = "longitude" ;
                lon:units = "degrees_east" ;
                lon:axis = "X" ;
        float tas(time, lat, lon) ;
                tas:units = "K" ;
                tas:code = 167 ;
                tas:table = 128 ;
                tas:cell_methods = "time: mean" ;
                tas:institution = "ECMWF" ;
                tas:standard_name = "air_temperature" ;
                tas:long_name = "Near-Surface Air Temperature" ;
                tas:coordinates = "time lat lon height" ;

// global attributes:
                :CDI = "Climate Data Interface version 1.9.8 (https://mpimet.mpg.de/cdi)" ;
                :institution = "European Centre for Medium-Range Weather Forecasts" ;
                :Conventions = "CF-1.6" ;
                :frequency = "mon" ;
                :CDO = "Climate Data Operators version 1.9.8 (https://mpimet.mpg.de/cdo)" ;
                :units = "hours since 2000-02-14 00:00:00" ;
                :_NCProperties = "version=2,netcdf=4.7.1,hdf5=1.10.5" ;
                :NCO = "netCDF Operators version 4.9.2 (Homepage = http://nco.sf.net, Code = http://github.com/nco/nco
```

Requesting data

- Gitlab project (private)
- To request data specifying variables, source, frequency, period, etc...
- or inform about potential issues, or questions regarding the organization
- No code
- 1626 requests in 4 years



requests

Project ID: 261

11 Commits · 1 Branch · 0 Tags · 829 KB Files · 1.3 MB Storage

Request petition system for Earth Department.

This project is only for creating issues to track requests. It does not contain code.

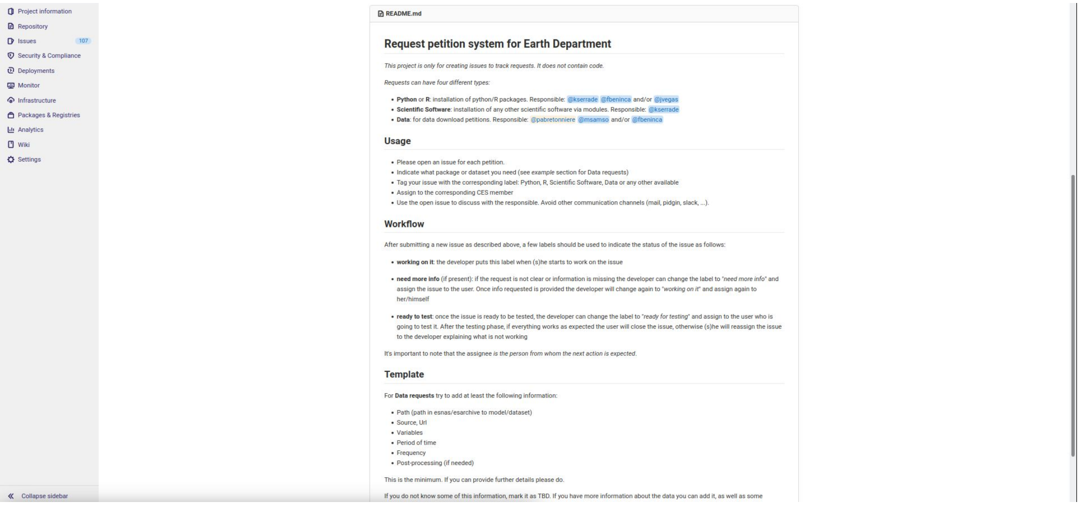
Auto DevOps

It will automatically build, test, and deploy your application based on a predefined CI/CD configuration.

Learn more in the Auto DevOps documentation

Backing up BSC Easybuild custom recipes
Kim Serradell authored 10 months ago

Name	Last commit	Last update
P9	Backing up BSC Easybuild custom recipes	10 months ago
WS/easybuild	Backing up BSC Easybuild custom recipes	10 months ago
nord3	Backing up BSC Easybuild custom recipes	10 months ago
README.md	Update README.md	1 year ago



Request petition system for Earth Department

This project is only for creating issues to track requests. It does not contain code.

Requests can have four different types:

- **Python** or **R**: installation of python/R packages. Responsible: @kserrade @fbeninca and/or @jvegas
- **Scientific Software**: installation of any other scientific software via modules. Responsible: @kserrade
- **Data**: for data download petitions. Responsible: @pabretonniere @msamso and/or @fbeninca

Usage

- Please open an issue for each petition.
- Indicate what package or dataset you need (see *example* section for Data requests)
- Tag your issue with the corresponding label: Python, R, Scientific Software, Data or any other available
- Assign to the corresponding CES member
- Use the open issue to discuss with the responsible. Avoid other communication channels (mail, pidgin, slack, ...).

Workflow

After submitting a new issue as described above, a few labels should be used to indicate the status of the issue as follows:

- **working on it**: the developer puts this label when (s)he starts to work on the issue
- **need more info** (if present): if the request is not clear or information is missing the developer can change the label to "*need more info*" and assign the issue to the user. Once info requested is provided the developer will change again to "*working on it*" and assign again to her/himself
- **ready to test**: once the issue is ready to be tested, the developer can change the label to "*ready for testing*" and assign to the user who is going to test it. After the testing phase, if everything works as expected the user will close the issue, otherwise (s)he will reassign the issue to the developer explaining what is not working

It's important to note that the assignee *is the person from whom the next action is expected*.

Template

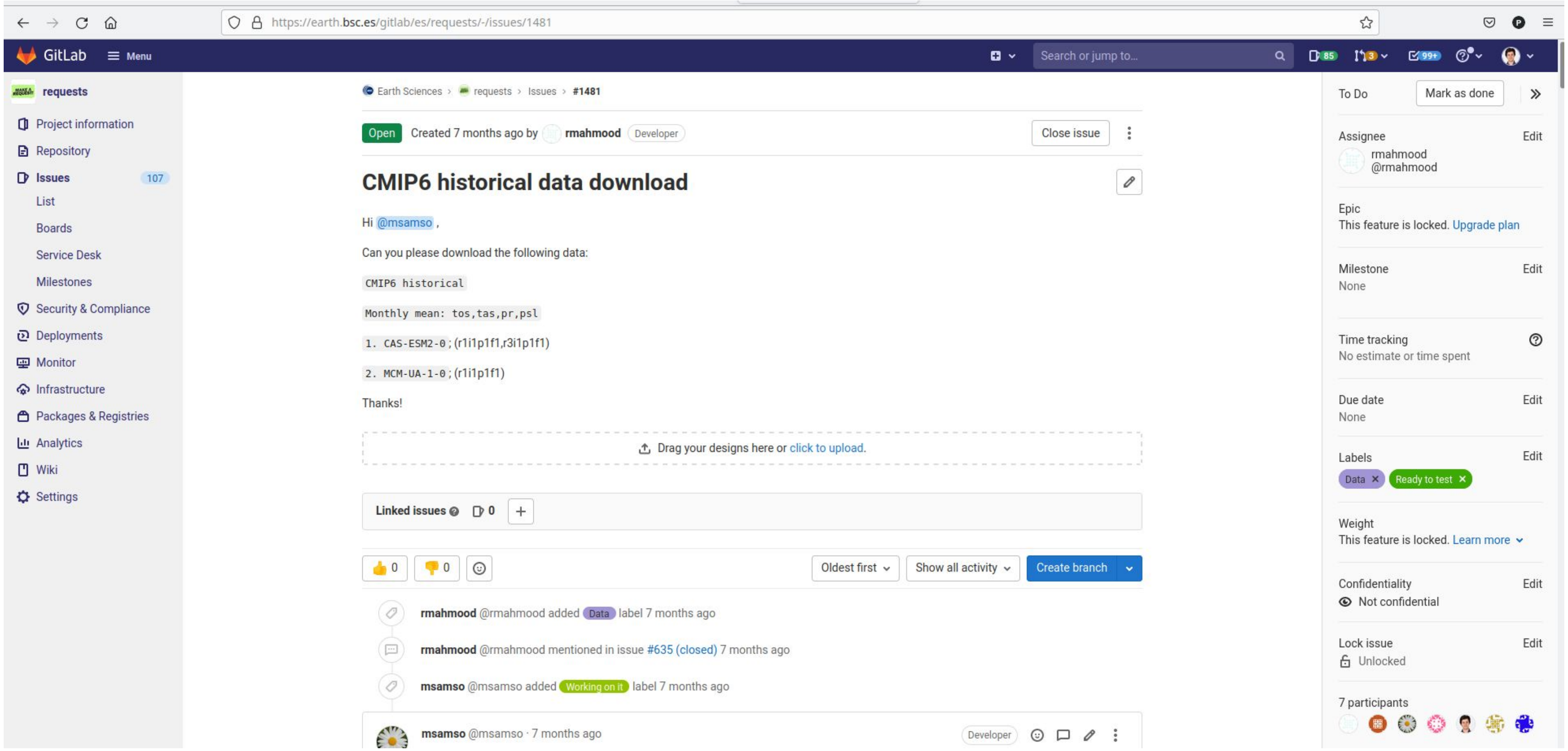
For **Data requests** try to add at least the following information:

- Path (path in esnas/esarchive to model/dataset)
- Source, Url
- Variables
- Period of time
- Frequency
- Post-processing (if needed)

This is the minimum. If you can provide further details please do.

If you do not know some of this information, mark it as TBD. If you have more information about the data you can add it, as well as some

Requesting data



Open Created 7 months ago by rmahmood Developer

CMIP6 historical data download

Hi @msamso ,

Can you please download the following data:

`CMIP6 historical`

`Monthly mean: tos,tas,pr,psl`

`1. CAS-ESM2-0` ; (r1i1p1f1,r3i1p1f1)

`2. MCM-UA-1-0` ; (r1i1p1f1)

Thanks!

rmahmood @rmahmood added Data label 7 months ago

rmahmood @rmahmood mentioned in issue #635 (closed) 7 months ago

msamso @msamso added Working on it label 7 months ago

msamso @msamso · 7 months ago Developer

Assignee: rmahmood @rmahmood

Epic: This feature is locked. Upgrade plan

Milestone: None

Time tracking: No estimate or time spent

Due date: None

Labels: Data, Ready to test

Weight: This feature is locked. Learn more

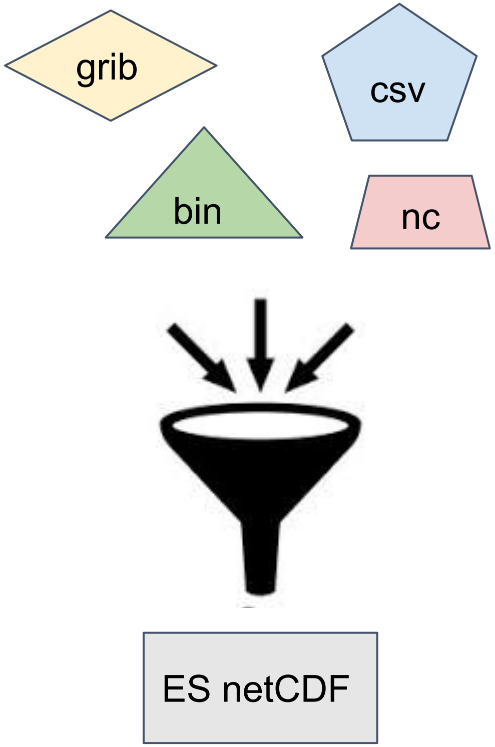
Confidentiality: Not confidential

Lock issue: Unlocked

7 participants

Data manipulation

- From heterogeneous sources, format and organization to department standards
- Format conversion
- File name and directory homogenization
- Adding/correcting metadata
- Scripts in bash or python, with packages, wget, aria2, ncftp, NCO, CDO, xarray, ...

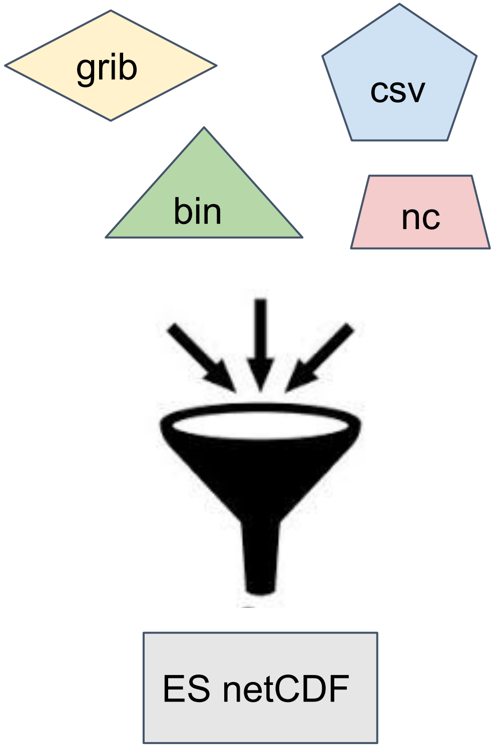


Data manipulation: generic tools examples

- CDS downloader: python tool to download era5 + seasonal forecast in netcdf based on a configuration file and cdsapi

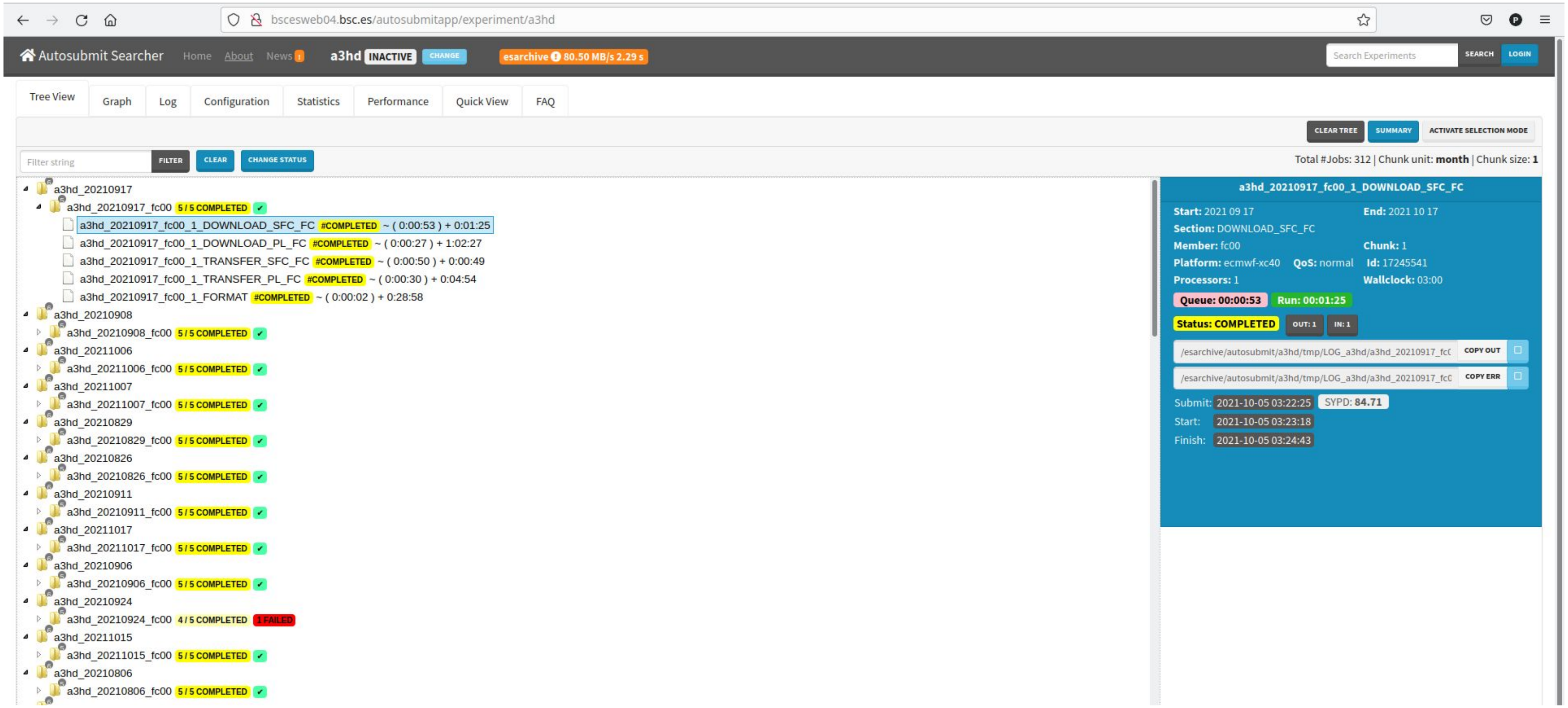
- R2D2/C3PO: python tool to manage ECMWF mars downloads request (grib) and formatting (conversion to nc, computation of means and modules, deaccumulations,...)

- ESGF sproket: download tool from ESGF



Data manipulation: generic tools examples

- auto-ESGF,auto-mars: extra layer on top of these tools with a workflow manager (Autosubmit) to orchestrate the parallelization of the different dates/variables/models/...



Total #Jobs: 312 | Chunk unit: **month** | Chunk size: **1**

```
a3hd_20210917
  a3hd_20210917_fc00 5 / 5 COMPLETED ✓
    a3hd_20210917_fc00_1_DOWNLOAD_SFC_FC #COMPLETED ~ ( 0:00:53 ) + 0:01:25
    a3hd_20210917_fc00_1_DOWNLOAD_PL_FC #COMPLETED ~ ( 0:00:27 ) + 1:02:27
    a3hd_20210917_fc00_1_TRANSFER_SFC_FC #COMPLETED ~ ( 0:00:50 ) + 0:00:49
    a3hd_20210917_fc00_1_TRANSFER_PL_FC #COMPLETED ~ ( 0:00:30 ) + 0:04:54
    a3hd_20210917_fc00_1_FORMAT #COMPLETED ~ ( 0:00:02 ) + 0:28:58
a3hd_20210908
  a3hd_20210908_fc00 5 / 5 COMPLETED ✓
a3hd_20211006
  a3hd_20211006_fc00 5 / 5 COMPLETED ✓
a3hd_20211007
  a3hd_20211007_fc00 5 / 5 COMPLETED ✓
a3hd_20210829
  a3hd_20210829_fc00 5 / 5 COMPLETED ✓
a3hd_20210826
  a3hd_20210826_fc00 5 / 5 COMPLETED ✓
a3hd_20210911
  a3hd_20210911_fc00 5 / 5 COMPLETED ✓
a3hd_20211017
  a3hd_20211017_fc00 5 / 5 COMPLETED ✓
a3hd_20210906
  a3hd_20210906_fc00 5 / 5 COMPLETED ✓
a3hd_20210924
  a3hd_20210924_fc00 4 / 5 COMPLETED 1 FAILED
a3hd_20211015
  a3hd_20211015_fc00 5 / 5 COMPLETED ✓
a3hd_20210806
  a3hd_20210806_fc00 5 / 5 COMPLETED ✓
```

a3hd_20210917_fc00_1_DOWNLOAD_SFC_FC

Start: 2021 09 17 **End:** 2021 10 17
Section: DOWNLOAD_SFC_FC
Member: fc00 **Chunk:** 1
Platform: ecmwf-xc40 **QoS:** normal **Id:** 17245541
Processors: 1 **Wallclock:** 03:00
Queue: 00:00:53 **Run: 00:01:25**
Status: COMPLETED OUT: 1 IN: 1
/esarchive/autosubmit/a3hd/tmp/LOG_a3hd/a3hd_20210917_fc(
/esarchive/autosubmit/a3hd/tmp/LOG_a3hd/a3hd_20210917_fc0
Submit: 2021-10-05 03:22:25 SYPD: **84.71**
Start: 2021-10-05 03:23:18
Finish: 2021-10-05 03:24:43

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Checkers

- standard compliance (CF checker + internal tools)
- time and space homogeneity (nctime + internal tools)
- visual inspection
- ESMValTool
- + some physical plausibility checks

Connection with tools

- Standardization is done to improve the data usability and discovery

But also for the tools:

- internal tools (verification, visualization, diagnostics) built to take advantage of this organization
- 2 ways interaction: conventions for the tools and tools following the conventions

Conclusions and main challenges

- Data organization is done following several levels of “strictness”
- Better data organization
    - less duplication
    - more use of the data
    - more efficient use
    - “findability”
    - improved efficiency of the tools
    - more external reusability
- Challenges with the tools and the checks
- Needs automation and centralization

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

Thank you
Questions?

pierre-antoine.bretonniere@bsc.es