

UPORTO Data Archive for Perdigao

If an institution wants to start uploading data to the archive, it must send an email to windscanner.perdigao@gmail.com requesting access credentials

This note explains the usage of the “UPORTO Data Archive for Perdigão” to store the datasets produced in the Perdigão experiment.

1. Introduction

The UPORTO Data Archive is one of the three data archives described in the “Operations Plan Chapter 2 | Perdigão Data Management Support” document¹:

Perdigão data will be made available to the community through designated distributed project archives at NCAR/EOL, DTU and the University of Porto (UPORTO).

The archive lives in a `ext4` file system managed by a Linux Ubuntu 16.04 LTS. RAID5¹ is used to achieve durability of data, as it provides good protection against and recovery from hardware defects.

Archived data is collected by a `rsync`² server running at UPORTO. Clients contact the `rsync` daemon directly via TCP which can be done periodically using a cron job or a windows task.

`Rsync` is a fast and extraordinarily versatile file copying tool. It is famous for its delta-transfer algorithm, which reduces the amount of data sent over the network by sending only the differences between the source files and the existing files in the destination.

2. How the UPORTO Data Archive works

The UPORTO Archive collects all data, raw and derived, produced in the Perdigao experiment, that passes the quality control procedures of the producer, as files of the archive.

Derived products — plots and charts — may also reside in the archive alongside the data.

The files may be later replaced by new versions with higher quality, if and when available, and files deleted in the sources will be automatically deleted in the archive (e.g. data is mirrored).

Any institution participating in Perdigão may use the UPORTO data archive as a security backup and to ensure durability of data.

3. How is data uploaded

Any institution producing data at Perdigão (“*producer*”) stores the data in UPORTO Archive by choosing a directory in its own file system (the “*origin*”) and start a job to copy its contents to an archive “*export*” (the “*target*”)³:

```
rsync -avz origin/ producer@windsptds.fe.up.pt::target
```

¹ https://en.wikipedia.org/wiki/Standard_RAID_levels#RAID_5

² <https://en.wikipedia.org/wiki/Rsync>

³ It is advisable to test the effect the command beforehand by using the option `--dry-run`

An export *target*, together with a password, is generated to be used only by the institution in its own data mirroring process.

The data available at the *origin* file system may be changed at any time and the archive will be automatically updated accordingly.

If the institution wants **UPORTO to deal with the mirroring setup** details, it suffices to provide UPorto an account with the rsync utility and with access to *origin* directory. UPorto will install the cron job running rsync every 4 hours of the day.

After the initial setup, the institution has only to manage the files in its file system which will be automatically copied to the UPORTO archive and eventually be replicated to the other two Perdigão data archives at DTU and UCAR.

4. Configuration details

In order to start the mirroring process, an institution (*producer*):

- Sends an email to windschanner.perdigao@gmail.com requesting access credentials (*producer* and *password*)
- Choose a suitable directory in its own file system (*origin*)

4.1. Using Unix machines

If the data lies in a **Linux** server machine, or in a **MacOS** machine, a rsync job does **an exact mirror**⁴, in the archive, of the content of the *origin* directory:

```
export RSYNC_PASSWORD=password
rsync -avz --delete origin/ producer@windsptds.fe.up.pt::export
```

To upload the *origin* directory to the root of the export:

```
rsync -avz origin producer@windsptds.fe.up.pt::export
```

To upload the *origin* directory inside a *destination* folder in the root of the export:

```
rsync -avz origin producer@windsptds.fe.up.pt::export/destination
```

4.2. Using Windows machines

In **Microsoft Windows** cwRsync is a free distribution of the rsync utility that is optimized for quick and easy installation. The producer may:

- download cwRsync: <https://itfix.net/content/cwrsync-free-edition>
- unzip to a suitable location (c:\archive)
- Create a data directory (mkdir c:\archive\data)
- run rsync.exe (substitute the password and the producer by the given credentials)⁵:

```
# goto the archive dir
cd c:\archive
# set the password
set RSYNC_PASSWORD=password
# mirror the content of your data directory
.\bin\rsync -avz --delete data/ producer@windsptds.fe.up.pt::producer
```

⁴ Behare of the `--delete` option effects; you may want to use `--dry-run` first to be sure

⁵ To use an absolute path, the windows path "c:\archive\data" is "/cygdrive/c/archive/data/"

To schedule the backup, which is done through Task Scheduler:

- open Task Scheduler by choosing: Start Menu, Programs, Accessories, System Tools, Scheduled Tasks.
- the Scheduled Tasks explorer will open. Double click the item named "Add Scheduled Task".

A. Crontab example

```
nejoco@login.neweuropeanwindatlas.eu> crontab -l
RSYNC_PASSWORD=*****
# At minute 31 past every 4th hour
31 */4 * * * /usr/bin/rsync -az --delete /newa/WP2/PERDIGAO/
dtu@193.136.38.237::dtu > /dev/null 2>&1
```

B. Current state of the Archive

An up-to-date detailed daily summary is available at: <http://winds.fe.up.pt/datalogs/?C=N;O=D>

B.1. Summary

```
> DTU: 1.8 TiB (4.08464 GiB)
> UCAR: 1.2 TiB (0 GiB)
> DLR: 1.2 TiB (0 GiB)
> INEGI: 301 MiB (0 GiB)
> WINDFORS: 3.2 GiB (0 GiB)
Summary on 20-11-2017 08:00.
```

B.2. Available exports

```
nejoco@VIND-pNEWA04:~> rsync -rdt rsync://windsptds.fe.up.pt
test                RSYNC test
archive             RSYNC UDA FILES (read only)
ucar                RSYNC UCAR FILES
dtu                 RSYNC DTU FILES
inegi               RSYNC INEGI FILES
dlr                 RSYNC DLR FILES
windfors            RSYNC WindForS FILES
```

B.3. Thredds data server

An experimental TDS, configured to serve the UPORTO Data Archive for Perdigão, is available at <https://windsptds.fe.up.pt/perdigao> (using the UCAR's FTP site credentials).

B.4. The catalogue

```
perdigao@windsptds:/data$ tree -d -L 3 /data/perdigao
```

```
data/perdigao
```

```
├── dlr
│   ├── HATPRO_level-1
│   │   ├── 201704
│   │   └── 201704_quicklook
│   │   ...
│   ├── HATPRO_level-2
│   │   ├── 201704
│   │   └── 201704_quicklook
│   │   ...
│   ├── HATPRO_surface-met
│   │   ├── 201704
│   │   └── 201704_quicklook
│   │   ...
│   ├── mcs_data
│   │   └── 20170430095732
│   │   ...
│   ├── netcdf_lidar
│   │   ├── DLR85
│   │   ├── DLR86
│   │   └── DLR89
│   └── raw_data
│       ├── DLR85
│       ├── DLR86
│       └── DLR89
├── dtu
│   ├── data
│   │   ├── DTU_Leica_Scanning
│   │   ├── DTU_Mast_Data
│   │   └── DTU_WindScanner
│   ├── docs
│   ├── photos
│   └── plots
│       └── DTU_WindScanner
├── inegi
│   ├── EnerconWindTurbine
│   ├── LeosphereWindcube
│   │   └── 01_RawData
│   └── LidarAerialSurvey_RawData
│       ├── Images
│       └── PointCloud
├── ucar
│   ├── isfs
│   │   ├── hr_noqc_geo
│   │   └── noqc_geo_notiltcor
│   ├── iss
│   │   ├── sodar-rass-site
│   │   ├── upper-orange-grove-site
│   │   └── west-profiler-site
│   └── ncas
└── windfors
```

|— 2017
| |— 201704
| |— 201705
| |— 201706
|— cross
|— 2017