How to: store and transfer your (PRECIOUS) data

Brought to you by the CAN

October 2023

Outline

All data are not equal

What to store and where

How to transfer

How to share

Outline

All data are not equal

What to store where

How to transfer

How to share

All data are not equal

For example, some data need to be shared while others need to be accessible only to one user or even encrypted.

- **documents**: small files
- **codes**: small files with complex history
- **experimental data**: small to huge files

All data are not equal

BUT all data need to be saved.

Outline

All data are not equal

What to store where

How to transfer

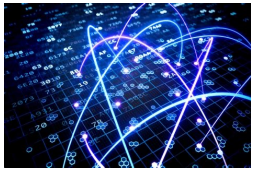
How to share

- **documents**: small files
- **codes**: small files with complex history
- **experimental data**: small to huge files

Where to store documents?

On an external device (hard disk). Possibly two backup copies, one away from the lab.

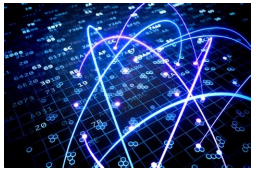
On the cloud

Nextcloud 

For ENS employees:

https://nextcloud.ens-lyon.fr

Nextcloud est un logiciel open source de partage et de synchronisation de fichier**.**

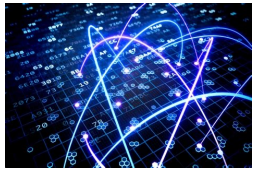
mycore 

For CNRS employee or members of a CNRS UMR:

https://ods.cnrs.fr/my_core.php

Up to 100 Go.

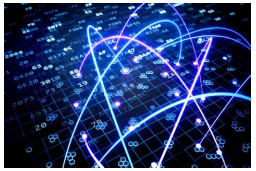
Can be shared with non-CNRS member

OSF 

https://osf.io

The OSF is a free open-source software project that facilitates open collaboration in science research.

Syncthing



https://syncthing.net/

Syncthing is a **continuous file synchronization** program. It synchronizes files between two or more computers in real time, safely protected from prying eyes.

- **documents**: small files
- **codes**: small files with complex history
- **experimental data**: small to huge files

Where to store and how to share codes

Use a distributed version control system

gitbio.ens-lyon.fr

You data is on a server with backup and on all the contributors computers

You can numerically sign and timestamp your contributions to the project

- **documents**: small files
- **codes**: small files with complex history
- **experimental data**: small to huge files

We can further categorize **experimental data** as :

- **hot**: data on which you are currently working on, you want a rapid access to them
- **warm**: data on which you may be working on, you want an easy access to them
- **cold**: data on which you will not be working on in a foreseeable future, you don't care if it takes some time to retrieve them.

The **hot** to **cold** categorization is closely related to the money and energy cost of the underlying storage facilities (the colder the cheaper).

The environmental impact is also a burning issue…

Backuped data versus **archived data**

Your data can have none to multiple **backups**

More **backups :**

- higher resilience
- higher costs

Backuped data versus **archived data**

Data that will not change in the future can be **archived**

Put your data in an archive facility along with the correct **metadata**

where it will get a unique identifier and will stay accessible *forever* (which may require a potentially large number of multi-site **backup**).

Data Management Plans (or **DMP**s): a key element of good data management.

A **DMP** describes the data management life cycle for the data to be collected, processed and/or generated.

It is an important par of making research data **FAIR** (findable, accessible, interoperable, reusable)

Beware: some funding agencies might require submission of formal DMPs

An online tool is available: https://dmptool.org

A **DMP** should include information on:

- the handling of research data during & after the end of the project
- what data will be collected, processed and/or generated
- which methodology & standards will be applied
- whether data will be shared/made open access and
- how data will be curated & preserved (including after the end of the project).

The **DMP** may need to be updated over the course of the project whenever significant changes arise.

Why on earth bother to write a DMP???


The first re-user of your data will be your future-self.
And past-self does not answer emails.
The first re-user of your data will be your future-self.
And past-self does not answer emails.
Florian @markowetzlab

Where to store experimental data

Whether local or worldwide, the solutions will depend upon the temperature of your data.

Biodata

The BIODATA storage space is managed by Stéphane Janczarski and hosted by the ENS DSI and allows to store raw data, directly from scientific platforms. Each ENS team has access to two folders:

Biodata 

•nameofteam/: (**2To** for all the LBMC), with daily snapshots on another server in the SING room

•nameofteam2/: (**12To** for all the LBMC), **backuped** monthly by Stéphane Janczarski

PSMN 

A copy of your data can be placed in your PSMN team folder /Xnfs/site/lbmcdb/team_name, with up to 600To of storage for the biology department.

This will also facilitate the access to your data for the people working on your project if they use the PSMN computing facilities.

PSMN 

The PSMN allows you to store hot to warm data BUT this is by no means an archive deposits space.

Thank's for a collaboration with IN2P3 (see below) it is straightforward to create a backup from the PSMN to the IN2P3

IN2P3

With a PSMN account, you can make long-term **backup** of your data there.

The CCIN2P3 don’t know you, and don’t provide archiving services, therefore you must write a **DMP** to define some information like the owner of the data, its nature and its lifetime.

IN2P3

You will need to create an account on dmp.opidor.fr, where you can find a **DMP** template for the CCIN2P3.

You can then contact the PSMN staff to send this **DMP** to the CCIN2P3. Once this **DMP** is validated by the CCIN2P3 staff, you will be able to upload your data from the PSMN.

IN2P3

The IN2P3 has develloped a tape-based archiving system called iRODS.

This is the privilege solution for long-term storage of large amounts of data

Public Archives

Public archives like ebi (UE) or ncbi (USA) are free to use for academic purpose

This is the « natural » way of archiving sequencing datasets

Public Archives

For example, for the ebi (UE) you have:

•ENA (the European Nucleotide Archive) to store raw sequencing data, sequence assembly information and functional annotation

•BIA (the BioImage Archive) to store and distributes biological images

Public Archives

Tip 1: plan some time ahead if you have never deposited in a public archive

Tip 2: public archives propose an **embargo** time system during which **your dataset will stay private**.

Public Archives

https://zenodo.org/

Max 50GB per dataset

- **Safe** — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- **Trusted** — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable** — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- **No waiting time** — Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- **Open or closed** — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning** — Easily update your dataset with our versioning feature.
- **GitHub integration** — Easily preserve your GitHub repository in Zenodo.
- **Usage statistics** — All uploads display standards compliant usage statistics

Outline

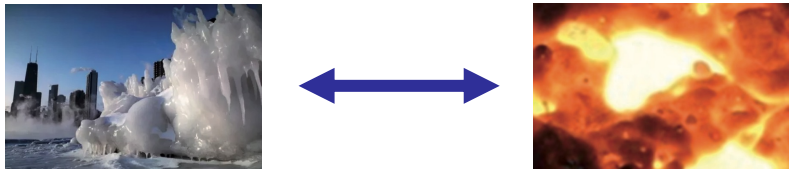
All data are not equal

What to store where

How to transfer

How to share

Transfert



The PSMN <-> IN2P3 connection has a 10Gbits/s speed.

Will depend upon the load of the PSMN server from which you transfer.

Example 1: 5.7TiB in 4 hours

Example 2: 1TiB in 3 hours

Thank's to Loïs Tautelle for those numbers

Outline

All data are not equal

What to store where

How to transfer

How to share

How to share documents?

Do NOT use commercial solution like Google Drive or Dropbox.

One of the cloud solution or an institutional git server

What the law says (in France)

LOI 2016-1321 du 7 octobre 2016 pour une République numérique

“Dès lors que les données issues d’une activité de recherche financée au moins pour moitié par des dotations de l’Etat, des collectivités territoriales, des établissements publics, des subventions d’agences de financement nationales ou par des fonds de l’Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu’elles ont été rendues publiques par le chercheur, l’établissement ou l’organisme de recherche, leur réutilisation est libre.”

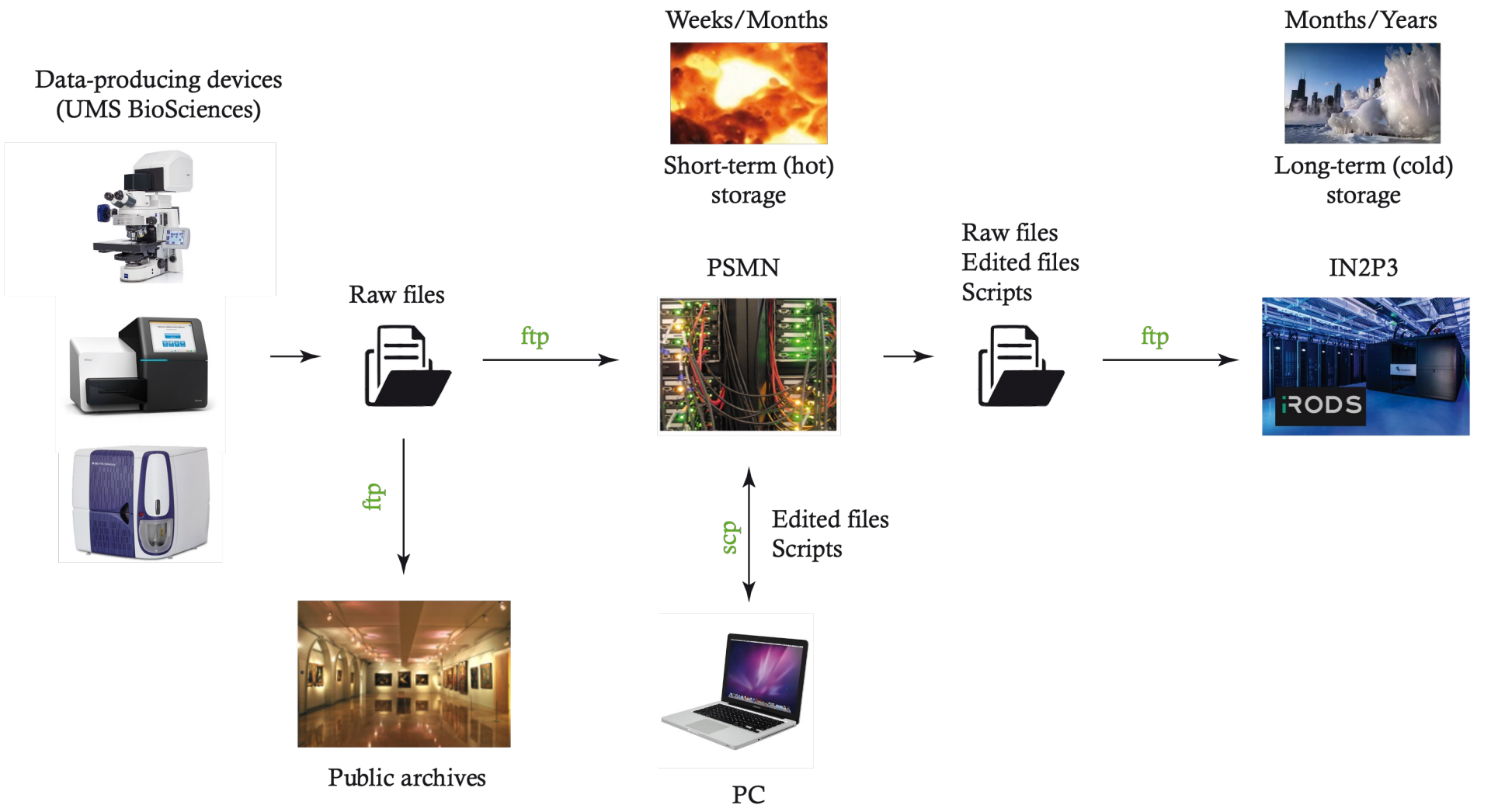
Free and open-source license

	Permissive	Weak Copyleft	Strong Copyleft
author's right/copyright guarantee	✓	✓	✓
use	✓	✓	✓
copy	✓	✓	✓
modification	✓	✓	✓
distribution	✓	✓(same license)	✓(same licence)
relicensing	✓	×	×
integration in non-copyleft software[18]	✓	≈ (c.f. next slide)	×

https://choosealicense.com/

https://creativecommons.org

A typical pipeline for data handling



Clinical data

Questions?