Data Storage and Open Repositories
Data sharing, metadata and data curation

Guillaume Anciaux

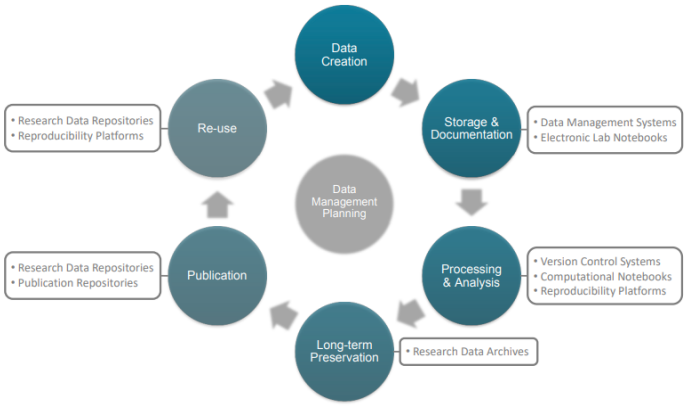
Guillaume Anciaux

- Senior computational scientist in mechanical engineering
- Mechanics and Numerical simulation
- Lecturer in Civil Engineering Institute
- Open (Data) Research
  - Member of ETH Domain Open Research Data (ORD) program
  - Data editor of the mechanics overlay Journal JTCAM
  - Developper of curation tools Solidipes

Purpose of the presentation

- Set some definitions
- List Open Data philosophy and incentives
- List storage possibilities (international and @EPFL,ETHZ, ...)
- Focus attention to FAIR principles
- Help you take the best decision for your data

Scientific data life cycle



Many of the following slides are based on a study conducted by the ETH Board ORD program aiming at reviewing the existing ORD services and infrastructures

Data

Etymology

https://www.etymonline.com/word/data

> 1640s, "**a fact given or granted**", classical plural of datum, from Latin datum "**(thing) given**", neuter past participle of dare "**to give**" (from PIE root *do- "to give").
>
> In classical use originally "**a fact given as the basis for calculation in mathematical problems.**" From 1897 as "**numerical facts collected for future reference.**"
>
> Meaning "**transmittable and storable information by which computer operations are performed**" is first recorded 1946.
>
> Data-processing is from 1954; data-base (also database) "**structured collection of data in a computer**" is by 1962;
>
> data-entry is by 1970.

Metadata

Definition

Metadata (or metainformation) is *"data that provides information about other data"*, but not the content of the data.

Summarizes basic information about data. Allows tracking and working with.

Types of metadata:

- **Descriptive metadata**: used for discovery and identification
- **Structural metadata**: how compound objects are organized (versions, directories, relationships)
- **Administrative metadata**: help manage a resource (type, standards, permissions, dates, locations)
- **Process data**: process used to create, analyse or transform the data
- **Legal metadata**: authorship, copyright, license

Metadata examples

Media type

A media type is:

- a two-part identifier for file formats and format contents
- somewhat similar to file extensions
- the Internet Assigned Numbers Authority (IANA) is the official authority for the standardization and publication of these classifications
- formerly known as a MIME (Multipurpose Internet Mail Extensions) type
    - denotes email message content/attachments type

Syntax notations

- a type and a subtype, which is further structured into a tree.

```
mime-type = type "/" [tree "."] subtype ["+" suffix]* [";" parameter];
```

Example (extracted from `/etc/mime.types`)

```
image/gif                                       gif
image/jpeg                                      jpeg jpg jpe jfif
```

Detection: `file` command on linux systems

```
file image.jpg
```

Image metadata:

- size of the image
- color depth
- resolution
- when it was created
- shutter speed.

=> make a live example with `exiftool` and/or `mediainfo`

- Metadata of images usually stored in the header of the file
- Different from the raw data pixel encoding

Text document's metadata

- how long is the document
- author
- when document was written
- a short summary

Web pages' metadata (Metatags)

- descriptions of page content
- keywords

Main factor for web searches (until the late 1990s)

"The reliance on metatags in web searches was decreased in the late 1990s because of "keyword stuffing", whereby metatags were being largely misused to trick search engines into thinking some websites had more relevance in the search than they really did."

Coma separated values

- Simple text format containing column values
- Header line contains the most important metadata (filename is another one)

HDF5

> Hierarchical Data Format (HDF) is a set of file formats (HDF4, HDF5) designed to store and organize large amounts of data.

- Provides lots of libraries in many languages to operate with such files

Open Documents formats

- Several formats for rich text (e.g. https://en.wikipedia.org/wiki/OpenDocument)
- What is metadata what is not ?
    - authorship ?
    - character fonts ?
    - language ?
    - Paragraphs ?

One way to decide: if I change format, **what data is totally preserved ?**

Relevance and usefulness of metadata

Without metadata a sequence of numbers could be:

- pixel image (data)
- signal measurements (data)
- file sizes (metadata)
- library indexes (metadata)

So should we use metadata to describe metadata ?

- Without context: **impossible to identify metadata**

Example:

- Context of filename and extension: accepted convention gives the context (and reading routines)

Metadata storage

metadata registry or metadata repository: metadata database

Example:

- File system: database of file metadata
- Image database software to treat a collection of images
- Any scientific repository (e.g. Zenodo) is also a database of metadata

Data storage

Definition

Data storage is the **recording** (storing) of **information** (data) in a **storage medium**

Examples

- Handwriting
- phonographic recording
- magnetic tape
- optical discs
- Biological molecules (RNA and DNA)

Fundamentally: recording is accomplished with energy (heat, electrical power, ...)

> Data storage in a digital, machine-readable medium is sometimes called digital data.

Measure

The **byte** is the unit of digital information, containing eight bits. Declined units are

Multiple-byte units V·T·E

Decimal			Binary			
Value	Metric		Value	IEC		Memory
1000	kB	kilobyte	1024	KiB	kibibyte	KB kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte	–
1000^6	EB	exabyte	1024^6	EiB	exbibyte	–
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte	–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte	–
1000^9	RB	ronnabyte			–	
1000^{10}	QB	quettabyte			–	

Orders of magnitude of data

Data entropy and compression

> In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes

Example:

- English is not as efficient as mathematics
- Files can be compressed (e.g. by analysing redundancy)

Data centers

Definition

> it a dedicated space used to house computer and network systems providing storage

Characterized by:

- redundancy or backup
- power supply (big energy consumer)
- data communication connections
- environmental controls (e.g., air conditioning, fire suppression)
- cyber-secure environment
- Duration: usually less than 10 years (put SCITAS cluster policy here). Mainly due to growth in produced data.

Annectode: https://www.datacenterdynamics.com/en/news/aws-building-site-burns-in-fatal-tokyo-fire-reports-say/

Datacenters maps

https://www.datacentermap.com/switzerland/

Modern Data centers

- **Standardization/consolidation**: allows to reduce number of data centers/and server sprawl
- **Virtualization**: servers can serve services (operating systems) on demand
- **Automating**: provisioning, configuration, patching, release management, and compliance
- **Securing**: Protection of virtual systems is integrated with the physical infrastructures

Data center design

Some of the considerations in the design of data centers are:

- **Size**: usually a building
- **Capacity**: can hold >= 1000 servers
- **Mechanical engineering infrastructure**: heating, ventilation and air conditioning (HVAC)
- **Electrical engineering infrastructure**: utility service planning; distribution, switching and bypass from power sources; uninterruptible power source (UPS) systems; etc...
- **Availability expectations**:

  > The costs of avoiding downtime should not exceed the cost of the downtime itself

- **Site selection**: proximity to power grids, telecommunications infrastructure, networking services, transportation lines and emergency services, geological risks, and climate (associated with cooling costs)
- **Overall costs optimization**: in dolars/CHF and CO2 emission...

Energy use

(More information here) Main article: IT energy management

Energy use is a central issue for data centers.

- from a few kW for a rack of servers in a closet
- up to several tens of MW for large facilities
- For higher power density facilities, electricity costs > 10% of the total cost of ownership (TCO) of a data center

Greenhouse gas emissions

> In 2020 data centers (excluding cryptocurrency mining) and data transmission each used about 1% of world electricity

This is huge, not even counting the CO2 emissions involved in production, buildings, etc....

Data centers Carbon footprint

Energy efficiency and overhead

Data center used energy efficiency metric:

- power usage effectiveness (PUE)
- power entering the data center / power used by IT equipments
- measures the overhead power: cooling, lighting, etc...), from 1.2 to 2.0
- European Union Code of Conduct for Data Centres

  > Calculations showed that in **two years** the **cost of powering and cooling** a server could be equal to **the cost of purchasing the server hardware**

- EPFL (SCITAS) changes computing/storage clusters every ~5 years
- EPFL tries to reuse energy: https://actu.epfl.ch/news/a-heating-plant-that-combines-renewable-energy-s-3/

Data center Tier standard

> Data center tiers are a standardized ranking system that indicates the reliability of data center infrastructure. This classification ranks facilities from **1 to 4**, with **1 being the worst** and **4 the best-performing level**.

Uptime Institute (independent organization) determines the Tier based on:

- Uptime guarantees
- Fault tolerance (redundancy, cooling and power infrastructure)
- Service cost
- Security levels
- Service speed (network, maintenance, client)
- Energy efficiency

Storage size elements of comparisons

- Google storage : above 20 exabytes = 20 million TB
- CERN: https://eos-web.web.cern.ch/eos-web/ : 700 PB = 700 000 TB
- EPFL total storage: ~20PB
- EPFL SCITAS computing clusters storage: ~10PB
- Personal computer: 512G - 4T
- A MP4 movie: 1G
- A MP3 song: 1M

Data repositories

Definition of repository

Etymology: late 15c. (Caxton), "**vessel**, for **storage**"

> From French *repositoire* or directly from Late Latin repositorium **"store"**

In classical Latin, **"a stand on which food is placed"**, from noun use of repositus, past participle of reponere **"put away, store"**

> The figurative sense of **"place where anything immaterial is thought of as stored"** is recorded from 1640s

Commercial sense of **"place where things are kept for sale"** is by 1759

Definition

> A receptacle or place where things are **deposited, stored, or offered for sale**
>
> An abundant source or supply; **storehouse**

What is the definition relevent to this presentation ?

Scientific definition: Archives and online databases

- **Content repository**: a database with an associated set of data management tools, allowing application-independent access to the content
- **Disciplinary repository**: an online archive containing works or data associated with a particular subject area
- **Academic document repository**: an archive where authors can deposit academic documents
- **Information repository**: a central place in which an aggregation of data is kept and maintained in an organized way, usually in computer storage
- **Institutional repository**: an archive for keeping digital copies of the intellectual output of an institution
- **Open-access repository**: a platform for freely available research results
- **Software repository**: a storage location for software sources or packages

@ETH: Data Acquisition, Storage and Annotation

- @ETH domain: basic storage service very common
    - NAS storage (with various Tiers)
    - long-term storage (on tapes)
    - and high-performance storage systems (e.g., parallel file systems on HPC clusters)
- Individual institutes or research groups frequently operate their own storage infrastructures
- Using these storage infrastructures requires skills and tools

Data File repositories

SWITCHdrive

- cloud storage service provided by SWITCH
- 100GB/user
- Owncloud

Polybox

- cloud storage service provided by ETHz
- 50 GB/user
- 234 TB in total

Private cloud storage services

- DropBox
- Google Drive (contracts with EPFL and ETHz)
- Microsoft OneDrive (contracts with ETHz)

> Public sharing of research data is relatively rare and does not conform to ORD and FAIR standards, as the

data is not discoverableFinally, publishing and sharing of very large datasets (TBs to PBs and upwards) remains a significant challenge in several research disciplines of the ETH Domain (e.g., climate science, microscopy).

Storage@EPFL

Data Storage: link

- Personal space: MyNAS (25GB, no sharing, no cost, redundency)
- Unit space: NAS WebDAV (1T)
- EPFL Object-S3 Storage Service (pay per use)

Remark: S3 access requires that all accessed files must be downloaded first: huge file sets may appear slow
Remark: Streaming solution, like JuiceFS project exist

Data sharing (sharepoint, google drive, switch drive)

- Non confidential: Google Suite
- Some documents are confidential:
  - Personal: SharePoint, SWITCHDrive (100 GB), MyNAS (25GB)
  - Team work: SharePoint, NAS WebDAV (1T)

> **Depending on the legual status of the project it may be important to guaranty storage in switzerland !**

Resources

- EPFL doc

Globus (High performance, large volumes)

- fast data transfer system
- managed by the University of Chicag
- efficient transfer of very large datasets

Within the ETH Domain Globus endpoints

- CSCS
- PSI
- ETHZ

Remark:

> For the data repositories of the ETH Domain, however, publishing and sharing of very large datasets

remains a significant challenge

Acquisition, Analysis,Treatment, Annotation@ETH

EPFL

- Slims: life-science oriented Laboratory Information Management System (LIMS) and Electronic Laboratory Notebook (ELN).
- eln.epfl.ch: ELN and repository for spectroscopic data
- RSpace: digital research platform for Institutional Research Data Management (free version, (paid) version being evaluated, not for every one)
- RedCap secure clinical science data management (service on-demand)
- Open Sample platform that allows scientists to search for antibodies, plasmids, cells or any other biomedical research tool. It allows a search on using the Research Resource Identifier (RRID, a universal identifier) and this gives information as to whether this material has been used at EPFL and the contact details of the associated laboratory.

ETHz

- openBIS: storage, annotation and backup of research data. (management solution, ELN, LIMS, web-based, efficient transfers, life-science, environment, materials). Interfaces with publication repositories (**ETH Research Collection** and **Zenodo**)

PSI

- Different ELNs (openBIS, Biovia ELN, ELOG and SciLog),
- Limsophy: LIMS as well as Data Catalog for raw and metadata capture at acquisition.

Quote from the report

> [...] the widespread adoption of ELNs in academic and industrial research has led to a huge

“zoo” of available ELN software solutions. However, the different **ELN solutions are not interoperable with each other** and this situation should be improved.

Resources

- active data management @ EPFL

Solution name	Main institution	Other institutions	Type	Domain	Selected statistics
Files & folders	All	N/A	Data management	All	This is the ‘default’ option in most research groups
openBIS	ETHZ	Empa, PSI, Eawag	Data management, ELN, LIMS	Quantitative Sciences	≈ 70 labs at ETHZ 13 labs at Empa Pilots at PSI & Eawag
Slims	EPFL	Unknown	ELN, LIMS	Life Sciences	≈ 70 labs at EPFL
Eln.epfl.ch	EPFL	None	ELN	Chemical Sciences	N/A
RSpace	EPFL	Unknown	ELN	Chemical Sciences	Users of free version unknown On-premise pilot in 4 labs
RedCap	EPFL	ETHZ	Secure data acquisition system	Clinical Sciences	N/A
Biovia ELN	PSI	None	ELN	Life Sciences	PSI BIO division
ELOG	PSI	Empa	Electronic Logbook		N/A
SciLog	PSI	None	ELN		N/A
Limsophy	PSI	None	LIMS	Scientificlaboratories	N/A

Software repositories and Version Control systems

> In version control systems, a repository is a data structure that stores metadata for a set of files or directory structure.

- **Distributed** (GIT-like): the whole information is replicated in every clone
- **Centralized** (SVN): the whole information only on a single server

Metadata

- a historical record of changes
- a set of commit objects
- a set of references to commit objects, called heads/tags/branches

Software/Code repository: Forges

> a forge is a web-based collaborative software platform for both developing and sharing computer applications

Remark: The term forge refers to a common names such as SourceForge stemming from the metalworking forges

It is a specific **data repository** allowing to:

- store sources, packages and applications
- communicate with coworkers
- track and solve bugs
- manage branching developments
- fork the project (create parallel developments)

Version Control @ ETH

For software sources and text-based data GIT is commonly used in the ETH Domain for:

- computational research (simulation, analytical workflows)
- code management and reproducible research

Platforms

- GitLab
  - deployed @EPFL, ETHZ, PSI, WSL, SDSC
  - Many local instances of gitlab
  - SWITCH Gitlab
- GitHub

Source code repositories

Source

- https://en.wikipedia.org/wiki/Repository_(version_control)
- https://en.wikipedia.org/wiki/Comparison_of_source-code-hosting_facilities#References

	Name	Free server?	Free client?
0	Assembla	No	Unknown
1	Azure DevOps Services	No	No
2	Bitbucket	No	No
3	CloudForge	No	Unknown
4	Gitea	Yes	Yes
5	GForge	Partial	Yes
6	GitHub	No	No
7	GitLab	Partial[9]	Yes[10]
8	GNU Savannah	Yes	Yes
9	Helix TeamHub	No	No
10	Launchpad	Yes	No
11	OSDN	Unknown	Yes
12	Ourproject.org	Yes	Yes

	Name	Web hosting	Code review
0	Assembla	Yes	Yes[21]
1	Azure DevOps Services	Yes	Yes
2	Bitbucket	Yes[24]	Yes[23]
3	Buddy	No	Yes
4	CloudForge	Yes	Unknown
5	GForge	Yes	Yes
6	Gitea	No	Yes
7	GitHub	Yes[30]	Yes[28]
8	GitLab	Yes[33]	Yes[32]
9	GNU Savannah	Yes	Yes[36]
10	Helix TeamHub	No	Yes[38]
11	Kallithea	Yes	Yes
12	Launchpad	No	Yes

	Name	Free server?	Free client?
13	OW2	No	No
14	Phabricator	Yes	Yes
15	SEUL	Unknown	No
16	SourceForge	Yes[17][18]	Yes

	Name	Web hosting	Code review
13	OSDN	Yes	Yes
14	Ourproject.org	Yes	Unknown
15	Phabricator	Yes	Yes
16	RhodeCode	Yes	Yes
17	SourceForge	Yes	Yes

	Name	CVS	Git	Hg	SVN	BZR	TFVC	Arch	Perforce	Fossil
0	Assembla	No	Yes	No	Yes	No	No	No	Yes	No
1	Azure DevOps Services	No	Yes	No	No	No	Yes	No	No	No
2	Bitbucket	No	Yes	Until Feb 2020[c]	No	No	No	No	No	No
3	Buddy	No	Yes	No	No	No	No	No	No	No
4	CloudForge	No	Yes	No	Yes	No	No	No	No	No
5	GForge	Yes	Yes	No	Yes	No	No	No	No	No
6	Gitea	No	Yes	No	No	No	No	No	No	No
7	GitHub	No	Yes	No	Partial[39]	No	No	No	No	No
8	GitLab	No	Yes	No	No	No	No	No	No	No
9	GNU Savannah	Yes	Yes	Yes	Yes	Yes[40]	No	Yes	No	No
10	Kallithea	No	Yes	Yes	No	No	No	No	No	No
11	Launchpad	Import only	Yes[14][41]	Import only[42]	Import only	Yes	No	No	No	Unknown
12	OSDN	Yes	Yes	Yes	Yes	Yes	No	No	Unknown	Unknown
13	Ourproject.org	Yes	No	No	Yes	No	No	No	Unknown	Unknown
14	OW2	Dropped[43]	Yes	No	Dropped[43]	No	No	No	No	No
15	Helix TeamHub	No	Yes	Yes	Yes	No	No	No	Yes	No
16	Phabricator	No	Yes	Yes	Yes	No	No	No	No	No
17	RhodeCode	No	Yes	Yes	Yes	No	No	No	No	No
18	SEUL.org	Yes	No	No	Yes	No	No	No	Unknown	Unknown
19	SourceForge	Dropped[44]	Yes	Yes	Yes	Dropped[45]	No	No	Unknown	No[46]

	Name	Users	Projects
0	GitHub	94000000	330,000,000[49]
1	GitLab	31190000	546,000[51][j]
2	Bitbucket	5000000	Unknown
3	Launchpad	3965288	40,881[54]
4	SourceForge	3700000	500,000[57]
5	GNU Savannah	93346	3,848[52]
6	OSDN	54826	6,294[55]
7	Ourproject.org	6353	1,846[56]
8	Assembla	-1	526,581+[47]
9	Buddy	-1	Unknown
10	CloudForge	-1	Unknown
11	Gitea	-1	Unknown
12	OW2	-1	Unknown
13	SEUL	-1	Unknown

Software package repositories

is a storage location for software packages.

- table of contents
- version handling
- repository managers

Package managers automatically install and update **packages** from repositories

Criterion	Package manager	Installer
Shipped with	Usually, the operating system	Each computer program
Location of installation information	One central installation database	It is entirely at the discretion of the installer. It could be a file within the app's folder, or among the operating system's files and folders. At best, they may register themselves with an uninstallers list without exposing installation information.
Scope of maintenance	Potentially all packages on the system	Only the product with which it was bundled
Developed by	One package manager vendor	Multiple installer vendors
Package format	A handful of well-known formats	There could be as many formats as the number of apps

Package formats

Each package manager relies on

- a format
- metadata
- group of files
- dependencies
- installation routines

Example: rpm, apt, dpkg (Linux), APK (Google), APPX, XAP (Microsoft Store)

	Metadata type	Used for
0	Versions available	Upgrading and downgrading automatically
1	Dependencies	Specify other artifacts that the current artif...
2	Downstream dependencies	Specify other artifacts that depend on the cur...
3	License	Legal compliance
4	Build date and time	Traceability
5	Documentation	Provide offline availability for contextual do...
6	Approval information	Traceability

	Metadata type	Used for
7	Metrics	Code coverage, compliance to rules, test results

	Language, purpose	Repository
0	Haskell	Hackage
1	Java	Maven[7]
2	Julia[8]	NaN
3	Common Lisp	Quicklisp[9]
4	.NET	NuGet[10]
5	Node.js	npm,[11] yarn, bower
6	Perl	CPAN
7	PHP	PECL, Packagist
8	Python	PyPI
9	R	CRAN[15]
10	Ruby	RubyGems[19]
11	Rust	crates.io[22]
12	Go	pkg.go.dev
13	Dart	pub.dev
14	D	dlang.org

	Package Manager	Description
0	npm	A package manager for Node.js[24]
1	pip	A package installer for Python[25]
2	apt	For managing Debian Packages[26]
3	Homebrew	A package installer for MacOS that allows one ...
4	vcpkg	A package manager for C and C++[28][29]
5	yum and dnf	Package manager for Fedora and Red Hat Enterpr...

Package Manager features/commands

	Action	apt	Homebrew	WinGet
0	Install package	apt install ${PKG}	brew install ${PKG}	winget install %PKG%
1	Remove package	apt remove ${PKG}	brew rm ${PKG} (rm is shorthand for remove or ...	winget uninstall %PKG%
2	Remove package (and orphans)	apt autoremove ${PKG}	brew rm ${PKG} && \ brew autoremove	winget uninstall %PKG%
3	Update software database	apt update	brew update	winget list > NUL
4	Show updatable packages	apt list --upgradable	brew outdated	winget upgrade
5	Delete orphans and config	apt autoremove	brew unlink ${PKG} && brew clean	—
6	Show orphans	—	—	—

Data Processing and Analysis

Computational Notebooks@ETH Domain for interactive scientific computing

- Jupyter notebooks, R Markdown
- Easy sharing of a “computational narrative” (code, documentation, results etc.)

Usage

- local computer
- JupyterHub
- Integration with [openBIS](https://openbis.ch/
- Online with RStudio, DeepNote, mybinder.org

ETH Examples

- @EPFL: JupyterHub Noto with SwitchAAI authentication
- @CSCS: JupyterHub platform for interactive computing on the supercomputer Piz Daint
- @WSL: JupyterHub connected to HPC resources
- Many ad-hoc JupyterHub servers in numerous research groups within the ETH Domain

Renku https://renkulab.io/

A framework initiative trying to lower the barrier-to-entry to current best-practices:

- developed by the Swiss Data Science Center (SDSC) as **OpenSource**
- **free service** (free storage!)
- GIT versioning
- Docker containers
- Datasets management
- Jupyter Notebook session on demand
- Knowledge Graph
- *Reproducibility*: Workflow management (on-going)
- S3 large space storage (on-going)
- Export/Import to other repositories (on-going)

Renku Instances

- Renku public (free) deployment https://renkulab.io: ~4000 users

    - 50-200 active interactive sessions at any given time
    - consumes approximately 1TB of RAM
    - 350 cpu cores
    - 20TB of object storage (data and container registry)
    - 40TB of active storage

- https://sv-renku.epfl.ch/ @ EPFL Life Science
- Instances at Fribourg/Lucern Universities

AiiDA https://www.aiida.net/

Supercomputer production of simulation data

- Costly to prepare run and store
- Crucial to keep track of results links

AiiDA@EPFL

- **open-source Python infrastructure**
- automating, managing, persisting, sharing and reproducing advanced simulation workflows
- support high-throughput computations (from seconds to weeks)
- interfaces with remote computation resources (job schedulers, data transport ...)
- Provenance graph natively maintained
- Interface with Renkulab
- Several plugins for many simulation codes, and workflows
- Jupyter-based interface **AiiDAlab**

Solution name	Institutions	Functionality	Users	Storage
GitLab	EPFL, ETHZ, PSI, WSL	Version control	Dependent on installation (e.g. 14'500 for ETHZ GitLab)	Dependent on installation (e.g. 6 TB for ETHZ GitLab)
JupyterHub	EPFL, ETHZ, CSCS, WSL, PSI	Interactive comp. notebooks	Dependent on installation	Dependent on installation
Renku/RenkuLab	SDSC, EPFL	Version control Interactive comp. notebooks Workflow management Provenance tracking Reproducibility	≈4000 for public instance	≈60 TB for public instance
AiiDA / AiiDAlab	EPFL, Empa, PSI	Interactive comp. notebooks Workflow management Provenance tracking Reproducibility	≈ 190	≈20 TB / year

Scientific repositories

FAIR principles

FAIR data meet principles of

- **F**indability
- **A**ccessibility
- **I**nteroperability
- **R**eusability

Remark: *The abbreviation FAIR/O data is sometimes used to indicate that the dataset or database in question complies with the FAIR principles and also carries an explicit data-capable open license.*

FAIR principles in details: GO FAIR

Findable

- F1. (Meta)data with unique and persistent identifier
- F2. Data richly described
- F3. Metadata includes the identifier of the data
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible

- A1. (Meta)data retrievable from identifier with standardised protocol
    - A1.1 Open, free and universally implementable protocol
    - A1.2 The protocol allows authentication (if necessary)
- A2. Always accessible Metadata (even if data not anymore available)

Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
    - R1.1. (Meta)data are released with a clear and accessible data usage license
    - R1.2. (Meta)data are associated with detailed provenance
    - R1.3. (Meta)data meet domain-relevant community standards

GO FAIR, FAIR Principles, https://www.go-fair.org/fair-principles/

Knowledge graph

A *knowledge graph* is a graph-structured data model that represents data.

Definition

A knowledge graph represents a network of real-world entities (objects, events, situations, or concepts)

This information is usually **stored in a graph database and visualized as a graph structure**

A knowledge graph is made up of:

- nodes (place, person, object, words, scientific output)
- edges (connection between nodes)
- labels (relationship type of the edge)

  prominently used by search engines (Google, Bing, ...)

For instance

- dataset **A** (node)
- paper **B** (node)
- A **is supplementing material for** B (edge and label)

Examples of knowledge graphs

- DBPedia
- Wikidata
- Google Knowledge Graph (represented through Google Search Engine Results, comprised of over > 500 million objects)
- Entertainment(Netflix): linking media based on viewers (also leveraged AI)
- Research: links between scientific data production (zenodo, renku, datacite)

Data repository example Wikidata

> collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation

It is a common source of open data that Wikimedia projects such as Wikipedia, and anyone else, can use under the CC0 public domain license.

an item (node) consists of:

- Obligatorily, an identifier (the QID), related to a label and a description.
- Optionally, multiple aliases and some number of statements (and their properties and values).

A statement (edge)

For example, the informal English statement "milk is white" would be encoded by

- a statement pairing the property color (P462)
- with the value white (Q23444)
- under the item milk (Q8495).

This is actually an edge.

> The most used property is cites work (P2860), which is used on more than 280,000,000 item pages as of January 2023

For knowledge (and scientific!) output, the edge and nodes labels are metadata!

Open and FAIR Scientific repositories

Zenodo

https://about.zenodo.org/policies/

> Zenodo is a *general-purpose open repository* developed under the European **OpenAIRE** program and operated by **CERN**

It allows researchers to deposit

- research papers
- data sets
- research software
- reports
- any other research related digital artefacts

A persistent **digital object identifier (DOI)** is associated, making the stored items easily citeable (and FAIR)

Size limitations (each entry < 50GB)

Zoo of Open repositories

- http://musam.imtlucca.it/wikisurf.html
- https://dataverse.harvard.edu/
- https://www.materialscloud.org/home
- https://openkim.org/
- https://arxiv.org/
- https://hal.science/
- others ...

Projects supporting Open Access Repositories and FAIR principles

DataCite

> international not-for-profit organization which aims to improve data citation

- easier access to research data on the Internet
- increase acceptance of legitimate research data
- citable contributions to the scholarly record
- support data archiving: verifyable&reusable results

How to establish easier access to research data

- assigning datasets a persistent identifiers (DOIs and others)

DataCite's recommended formats for a data citation is:

- Creator (PublicationYear): Title. Publisher. Identifier
- Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier

Remark:

Third-party tools allow the migration of content to and from other services such as ODIN, for ORCID

Founded in 2009 by organisations from 6 countries:

- British Library;
- Technical Information Center of Denmark (DTIC);
- TU Delft Library from the Netherlands;
- National Research Council's Canada Institute for Scientific and Technical Information (NRC-CISTI);
- California Digital Library (University of California Curation Center - UC3);
- Purdue University (USA);
- German National Library of Science and Technology (TIB).

February 2010: five additional members

- Australian National Data Service (ANDS);
- Deutsche Zentralbibliothek für Medizin (ZB MED);
- GESIS – Leibniz Institute for the Social Sciences;
- French Institute for Scientific and Technical Information (INIST);
- Eidgenössische Technische Hochschule (ETH) Zürich.

DataCite is an association and official DOI registration agency within the International DOI Foundation (IDF).

OpenAIRE

> Non-Profit Partnership, established in 2018 to ensure a permanent open scholarly communication infrastructure to support European research.

Vision

- transform society through validated scientific knowledge
- allow citizens, educators, funders, civil servants and industry to make science useful

Mission

- **shift** scholarly communication towards **openness** and **transparency**
- **facilitate** innovative ways to **communicate** and **monitor** research.

How?

- **Align policies**: network of 37 National Open Access Desks (NOADs).
- **Facilitate interoperability**: guidelines and services
- **Link research**: discoverability, transparency, reproducibility
- **Monitor Open Science**: work on science metrics
- **Train for Open Science**: researchers, policy makers, citizens...

In practice?

- participatory infrastructure with open governance
- Services developed by the community and for the community
- Empowering multiple and diverse Open Science Communities
- Respecting equitable and inclusive representation
- Seeking innovation on services, but no lock-in on content.
- Building European and global partnerships.

It also

- is a Network of Open Access repositories, archives and journals that support Open Access policies
- supports the implementation of the EC and ERC Open Access policies

> The **Zenodo research data repository** is a product of **OpenAIRE**

Registry of Research Data Repositories (re3data.org)

> Open science tool offering an overview of existing international repositories for research data

- All academic disciplines
- List existing research data repositories
- Help researchers to identify a suitable repository
  - (allows to comply with data policies requirements)
- Officially launched in May 2013

Content

- 2361 research data repositories (in July 2019)
- worldwide
- all academic disciplines
- described in detail (using the re3data.org schema)
- all metadata in the registry available for open use under the Creative Commons deed CC0

Inclusion criteria

The repository has to

- Focus on research data
- Be run by a legal entity (e.g. library, university)
- Clearly state access conditions to the data and repository
- Declare terms of use
- Show an English graphical user interface (GUI)

Context

> re3data.org is now a regular service of DataCite.

Let's try it https://re3data.org

Registry of Open Access Repositories

openROAR: http://roar.eprints.org/

> *The Registry of Open Access Repositories (ROAR)* is a searchable international database indexing the creation, location and growth of open access institutional repositories **and their contents***

openDOAR: https://v2.sherpa.ac.uk/opendoar/

> Directory of Open Access Repositories is a (UK-based) website listing open access repositories (including academic ones).

Remark openDOAR

- does not require **complete repository details** and does not search **repositories metadata**
- OpenDOAR **controls** submission of materials and is dependent on **the discretion of its staff**

Let's try them (fracture keyword)

Data Repositories within @ETH domain

Data Publication and Reuse

How to make Data Reusable ?

- public identifier
- accessibility

For small to medium size datasets (e.g., up to 100s of GB), this is typically achieved by publication in a **data repository**, with data hosted:

- in institutions (< 50 to 100GB)
- in general repositories (Zenodo < 50GB)

Repository name	Hosting institution	Other data-providing institutions3	Repository type	Software	PID type	Selected statistics4
Digital Object Repository at the Four Research Institutes (DORA)	Lib4RI	Eawag, Empa, PSI, WSL	Publication	Islandora	DOI	≈ 75'600 publications
Infoscience	EPFL	None	Publication	Invenio	DOI	≈ 162'000 publications
Eawag Research Data Collection (ERIC)	Eawag	None	General data	CKAN	DOI	≈ 150 open datasets ≈ 500 internal datasets
ETH Research Collection	ETHZ	None	Publication General data	DSpace	DOI	≈ 241'000 publications ≈1'700 datasets (42 TB total volume)
Data Catalog	PSI	Facility users, CSCS (see remarks)	General data	SciCat	DOI, PID	> 400'000 datasets, 9 PB total volume, >1'600 groups of users
Zenodo	CERN	Open	General data	Invenio	DOI	
EnviDat	WSL	Collaborations approved by WSL (see remarks)	Domain-specific (Environmental Sciences)	CKAN	DOI	≈ 540 datasets (20 TB total volume)
Materials Cloud	EPFL	PSI,Empa	Domain-specific data (Materials Sciences)	Invenio (customized)	DOI	≈ 22M crystal structures ≈ 7.5M simulations
Living Archives	EPFL	None	Domain-specific data (Architecture)	In-house	PID	≈ 11'000 items

Publication Repositories with @ETH domain

Storage

- working papers
- journal articles
- doctoral theses
- presentations
- supplementary materials

platform DORA 4RI

- Digital Object Repository at Four Research Institutes (Eawag, Empa, PSI, WSL)
- institutional repository and bibliography for all research articles/publications
- open-source software framework Islandora

Infoscience

- EPFL institutional repository, maintained by EPFL Library
- Currently: open-source software Invenio
- Migrate to: open-source software Dspace CRIS provided by 4Science.
- A Digital Object Identifier (DOI) is assigned on demand

ETHz Research Collection

- operated by the ETH Library as institutional repository for publications and

research data at ETH Zurich

- The Research Collection is based on the open-source software DSpace

Research Data Repositories

Research data repositories (in the following list) can be classified as general or domain-specific data repositories
General purpose repositories accept research outputs of all types without subject-specific focus.

ETHz

- the *Research Collection* is an institutional general-purpose data repository (ETH Library)
  - 1700 data items, 42 TB in total
  - Assign Digital Object Identifiers (DOI)
  - Data access rights (including embargo&retention)
  - Content preview (ZIP-files)
  - Connection with **ETH Data Archive** and external systems e.g. openBIS
  - Restriction on file upload sizes

Discussion

> *It can safely be assumed that usage of the Research Collection for research data will grow considerably over the next years as more researchers get used to making their data available.*
>
> *A major advantage of the Research Collection is its integration with numerous systems inside and outside of ETH Zurich. This level of integration prevents a shared use of the system together with other institutions. To achieve the same level of integration, institutions would need their own installations.***

EPFL

- Zenodo is the generic research data repository most widely used by EPFL researchers.
  - operated by CERN and OpenAire, with data stored in the CERN Data Centre.
  - the use of Zenodo for long-term retention and findability of code is suggested by Github
  - A Zenodo Curation policy has been added to the EPFL Community
- SDSC's Renkulab platform: data publication as Renku datasets
  - datasets are fully searchable
  - reusable across different projects.
  - No DOI but can be imported/exported from/to external repositories (e.g. Zenodo, OLOS and Dataverse)

Eawag

- Research Data Institutional Collection (ERIC)
- ERIC is comprised of two distinct parts:
  - an internal repository: ERIC-internal,
  - an open data portal: ERIC/open.
- in accordance with Eawag's data privacy guidelines
- ERIC-internal repository: data preserved
  - at least for scientific integrity requirements
  - at perpetuity sometime: e.g. environmental time-series
- ERIC/open data registered with a DOI (obtained from DataCite)
- EAWAG member of Datacite
- ~19Tb

WSL

Environmental Data portal (EnviDat).

- Specialized on environmental research data
- Hosts environmental research data (CH+World) published with DOIs
- Data provided by WSL, as well as EPFL, ETHZ, Eawag, PSI. (limited to environmental data)
- possibility to restrict access (embargo periods)
- based on CKAN
- half of the datasets qualify for opendata.swiss
- ~20TB

PSI

- Scicat Data Catalog
- Opensource SciCat software (collaboration with the European Spallation Source and the Swedish MAXIV synchrotron)
- unique persistent identifier
- published datasets assigned a DOI
- data tagged with searchable metadata
- Petabyte Archive System
- tape-based longterm storage system located at CSCS

Domain-specific data repositories

- Materials Cloud (Material csience)
  - seamless sharing and dissemination of resources in computational materials science
  - curated high-quality data (Materials Cloud Discover section)
  - OPTIMADE REST API
  - Materials Cloud Archive hosting worldwide materials science datasets, DOIs assigned to each of them
  - Reommended as a material science repository: SNSF, EU Commission, Nature
  - 22 million crystal structures, 7.5 million of which have associated density-functional-theory simulations, and with over 650'000

fully-reproducible simulations managed with AiiDA.
    - integrated with SDSC's RenkuLab: allows inspecting the data directly inside RenkuLab

- EPFL Living Archives (Architecture)
    - Architecture in collaboration with ENACIT4Research, harvesting Infoscience, to promote information and knowledge sharing via an online tool
    - facilitate access to the outputs from the Architecture department
    - organizes research and students works according to open tag structure

Persistent Identifiers

Persistent identifiers (PIDs) are long-lasting references to digital resources.

- essential to reliably identify datasets (for example in repositories)
- identify a data artifact independently from the physical place where it is located
- Persistent identifier allows translation to the data object's current location
- Most common one: Digital Object Identifiers (DOIs) are currently used for this purpose

ETH Zurich DOI Desk

- https://library.ethz.ch/en/researching-and-publishing/publishing-and-registering/doi-desk.html
- ETH Library
- Datacite
- registers DOIs for primary data (research data) and for secondary data such as working papers, articles or doctoral theses.
- available to all organizational units from Swiss higher education and research institutions but not to individuals.
- integrated with DataCite's global infrastructure
- ETH Zurich's IT Services provide technical infrastructure and support.
- DOI registration is free for all ETH Zurich organizational units.
- Other academic institutions and non-profit organizations must pay the fees defined by DataCite.
- 3'200'000 DOIs registered until December 2022
- The DOI Desk currently provides the registration service to 77 institutional customers, of which 50 are

not part of ETHZ (for example for WSL's EnviDat, Materials Cloud, or EPFL InfoScience).

> DOIs may not be appropriate as identifiers for all kinds of research data. For this reason, alternative PID

systems have been created

EPIC Identifiers

- eResearch Persistent Identifier Consortium (ePIC)
- CSCS provides and resolves ePIC PIDs for academic institutions in Switzerland
- The ePIC consortium offers a service to create, manage, and resolve persistent identifiers
- unique PID with a high degree of flexibility and robustness

SoftWare Heritage persistent IDentifiers (SHWID)

- Specific to software permanent entries (see below)

Data curation

Definition

Data curation is the **organization** and **integration** of data collected from various sources.

It aims at maintaining **the value of the data**, its **availability for consultation** and **re-use**, over time.

> In broad terms, curation means a range of activities and processes done to **create**, **manage**, **maintain**, and **validate**

> Specifically, data curation is the attempt to determine what information is worth **saving** and for **how long**

Curation allows **annotation (metadata)**, and allows **publication**, **presentation** and **preservation**

Remark: Datacite metadata schemas or Zenodo metadata description describe a norm to metadata describing a standard dataset

Scientific Journals policies

Policies of journals concerning data retention varies a lot with the discipline

- *Astronomy* field much better than *mechanical engineering*

One study:

B. Jackson. Open Data Policies among Library and Information Science Journals

> Open access journals in the discipline have disproportionately adopted detailed, strict open data policies
>
> Commercial publishers, which account for the largest share of publishing in the discipline, have largely adopted weaker policies.
>
> Rigorous policies, adopted by a minority of journals, describe the rationale, application, and expectations for open research data, while most journals that provide guidance on the matter use hesitant and vague language.
>
> Recommendations are provided for strengthening journal open data policies.

JTCAM curation example

> The Journal of Theoretical, Computational and Applied Mechanics is a scholarly journal, provided on a Fair Open Access basis, without cost to both readers and authors. The Journal aims to select publications of the highest scientific calibre in the form of either original research papers or reviews.

- Overlay journal
- Open review
- Experimenting datasets review
  - Curation help for the authors

Example JTCAM curated data set

https://renkulab.io/projects/guillaume.anciaux/jtcam-data-9733

https://zenodo.org/record/7729452

- Respect the Zenodo metadata documentation
- Respect the file format standards (when applicable, e.g. CSV convention)

Brain Imaging Data Structure

> The Brain Imaging Data Structure (BIDS) is a simple and intuitive way to organize and describe data.

- This is an example of strictly defined structure of files and metadata
- Allows to develop tools to help users create and read dataset (validating, reading and extraction tools)

Data Preservation and Disposal

> **Long-term preservation** ensures that **valuable research data**, and by extension **scientific results**, remain **interpretable and reusable** for years to come.

curated vs. non-curated

- Curated solutions (archives): typically in the library domain
- Non-curated long-term storage: does not ensure interpretability of data
  - File formats may no longer be readable
  - Data may be poorly described or not at all
  - Data missing anotations (or with inaccessible metadata) shall become inaccessible
  - Missing knowledge graph

Curated long-term preservation solutions in the ETH Domain

- EnviDat (WSL) and Data Catalog (PSI) implement long-term preservation of published data
- The ETH Data Archive (For ETHz members)
  - Storage repository with long term retention capabilities (10 years)
  - Compatible with the OAIS standard
  - Operated by the ETH Library
  - Mainly open-source **source code packages** or other **Research Data items**

Remark: provides guidelines to curation/publicatoin https://documentation.library.ethz.ch/display/DD/Instructions+and+fact+sheets

- EPFL Academic Output Archive (ACOUA, for EPFL members)
  - Storage repository with long term retention capabilities (10 years)
  - Compatible with the OAIS standard
  - Curated by the EPFL Library
  - Data is securely archived with metadata
  - Provide persistent identifiers
  - Capability to export datasets to Zenodo

Remark: Sadly the service has no success and maybe stopped

- CSCS Long Term Storage (LTS)
  - For CSCS users
  - Provides persistent identifier
  - Storage repository with long term retention capabilities (10 years)
  - persistent identifiers
  - Ability to set public access to data when needed
  - Data stored in LTS easily accessible from a web browser (HTTP protocol)
  - RESTful API to integrate with third party applications/portals
  - Scalable service that can cope with large volumes of data
  - Resiliency due to data protection measures against hardware/software failures
  - Clear licensing of the data

For most of them: some principles of the FAIR quadrant are not always included

Public copyright licences

> A public license or public copyright licenses is a license by which a copyright holder as licensor can grant additional copyright permissions to any and all persons in the general public as licensees

Usually, copyright holders give permission for others to

- copy
- modify
- reuse

Creative Commons CC0 1.0 Universal

- free access, and freedom to use the work as you wish ("use" includes to run a program or to execute a music score)

- freedom to access the "source-code" and use it as you wish, for study or change it for personal use
- freedom to redistribute copies
- right to quote (freedom to redistribute copies of fragments)
- freedom to distribute copies of your modified versions to others

GNU GPL for software, with similar guaranties

Remark: there could be law restrictions in some countries

Software preservation, Workflows and Containerization

Github/**Gitlab** repositories are volatile: nothing presents someone to rewrite the history, delete branches or all sorts of alteration of the metadata

Better to save a copy of a useful version (used to produce some data) as a "Dataset" (for instance in Zenodo)

Softare heritage

> Software Heritage (SH) provides a service for archiving and referencing historical and contemporary software

- Software repositories with retention guaranties
- SH was unveiled in 2016 by **Inria** and is supported by **UNESCO**
- non-profit multi-stakeholder initiative
- SoftWare Heritage persistent IDentifiers (SHWID): dedicated permanent identifier

Workflow preservation

> A workflow consists of a series of steps, each of which may involve

- input files
- code/software/application
- production of output files

Remark:

> The outputs of one step are frequently the inputs of another

- Creates a connection (knowledge graph sense) between the code execution and results
- For more complex workflows, the book-keeping can be tedious

Usage of **Renku**, **GNU-Makefile**, **CMake**, **snakemake**, **AiiDA** or **BlackDynamite** software may help for this (with more or less user involvements in describing the workflow)

Containerization

Within a long retention context, **the evolution of software versions** may become a problem

> **Containerization** is an increasingly popular way to address this issue.

Definition:

> In software engineering, containerization is operating system-level virtualization or application-level virtualization allowing software applications to run in **isolated and controlled** spaces called **containers**

Allows to pick a specific:

- Operating system version
- Software stack
- User and permission environment

Remark: This allows to choose/keep track of the exact version of software employed at the processing/analysis/creation of the data.

Remark: the containerization technology has been widely adopted by cloud computing platforms to ease deployment of application. Yet it suits very will reproducibility requirements

Docker

- set of platform as a service (PaaS) products
- use OS-level virtualization
- a **Docker Engine** hosts and manages the containers

Registries: A Docker registry is a repository for Docker images, ready to be deployed as containers

https://hub.docker.com/

Usage

- Needs a **Docker image description (Dockerfile)**: instructions to install necessary packages
- e.g. JTCAM dataset example using Renku templates

Kubernetes

Kubernetes is a famous open-source container orchestration system

- automating software deployment
- scaling
- and management
- designed by Google (now maintained by the Cloud Native Computing Foundation)

Remark: The name Kubernetes originates from Ancient Greek, meaning 'helmsman' or 'pilot'. Kubernetes is often abbreviated as K8s, counting the eight letters between the K and the s (a numeronym)

*Remark: used internally for https://noto.epfl.ch, https://gitlab.com and many web services

Remark: deployment of a renku instance requires a K8s deployment

Data curation: what/when to keep data ?

The data curation process is essential to **minimize** the size of dataset while keeping **relevent** data

- Compression
- Cleaning
- Annotation (non labelled data shall become useless sooner or later)
  - Links with other data (papers, peer work, etc)
  - Allows FAIR principles
- Consider ecology matters

Open discussion

> If the cost to reproduce the (simulation) data is lower than the cost to store it over a targeted period (that can amount to year), storing the minimal information to reproduce the data maybe enough

Example

- **Raw Image data** should be (zero-loss) compressed, annotated and kept
- The **input files** used to produce an analysis should always be kept
- The source software should be kept under **a permament id (SWHID)**
- The software context maybe stored within a **container**
- **Large analysis sets** produced by **analysis** should be kept only if
  - The computation is long (and costly) to produce
  - The access made by the scientific community is very regular
  - Otherwise it can be considered to discard it ?

Useful References

- Practical guide to submit datasets to Zenodo

https://www.epfl.ch/campus/library/wp-content/uploads/2018/09/DataDeposition_GoodBadUgly.pdf

- General official EPFL recommendations

https://www.epfl.ch/campus/library/services-researchers/data-publication/

- EPFL curation policy

https://zenodo.org/communities/epfl/about/

- Zenodo metadata documentation

https://github.com/zenodo/developers.zenodo.org/blob/master/source/includes/resources/deposit/_representation.md#deposit-metadata

- Tool put in place by library to list EPFL repositories

https://www.epfl.ch/campus/library/services-researchers/data-publication/data-repositories-and-related-platforms/

Conclusion

Preservation principles are possible with various

- time span
- data (and metadata) types
- repositories
- interactive/sharing possibilities
- level of curation

> ... several dedicated ORD solutions have been developed in the ETH Domain and are now in operation and recognized

both on a national and international level

> Nevertheless, the landscape of services and infrastructures is rather fragmented.

@EPFL the long term curated strategy is currently **Zenodo**

Remark: this is a difficult, on-going effort, sometimes not very well documented, making decisions as a researcher difficult

But if curation is made **with FAIR principles in mind**, the choice for the actual repository storing the data maybe non-important (DOI-like permanent identifiers can point to changing locations)