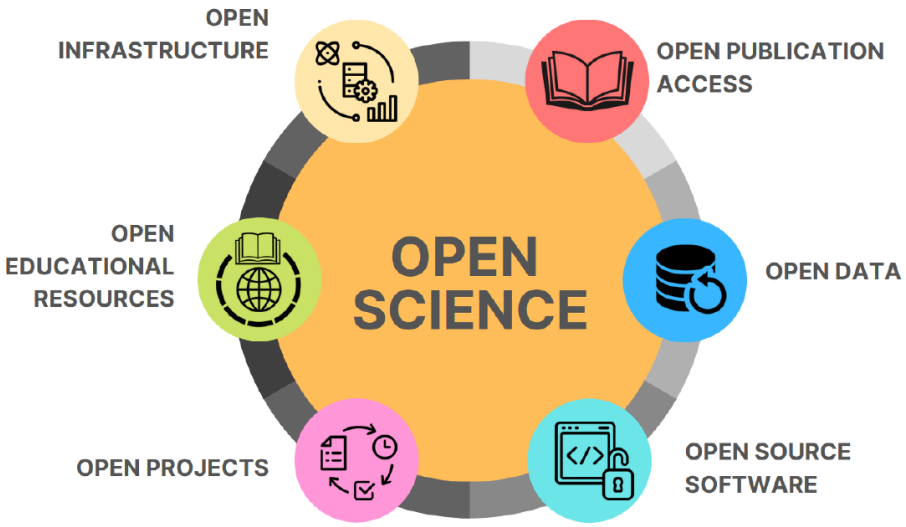
Introduction to Open Science and (FAIR) Data

A. Lemasson
several slides of this presentation have been adapted from
A. Mistry, O. Stezowsky, X. Espinal

What is Open Science?

- Science lives on the **open** exchange of knowledge
- ***Openness*** -> offers new prospects in the entire scientific research cycle and ***enables research outputs to be made openly accessible*** *and* ***broadly reusable*** (in sustainable infrastructures).
- This ***open culture*** of scientific endeavor is captured by the term "***Open Science***," Defined by the concepts of: ***transparency, sustainability, transfer, collaboration and sharing***
- Making **research outputs** (+ infrastructure) **publicly** available to science, industry, and society for **reuse** with ***as few barriers as possible***
- How to define and shape Open Science practices, tools and dissemination in a way that maximises the rewards and benefits? Important to consider what is ***useful for researchers!***
- **Open Science requires a shift in research culture**: it takes additional work and resources to practice open science: **Support needed from leadership**



Why is Open Science important ?

Accelerates **knowledge transfer** by breaking down access barriers to research outputs.

Fosters **collaboration** within and across disciplines, leading to quality improvements and new solutions.

Open Science promotes **transparency**, building **trust** among researchers and the public.

Attracting **future researchers**: Open Science signals inclusivity and appeals to diverse talent.

Sustainability improves as resource sharing reduces repetitiveness

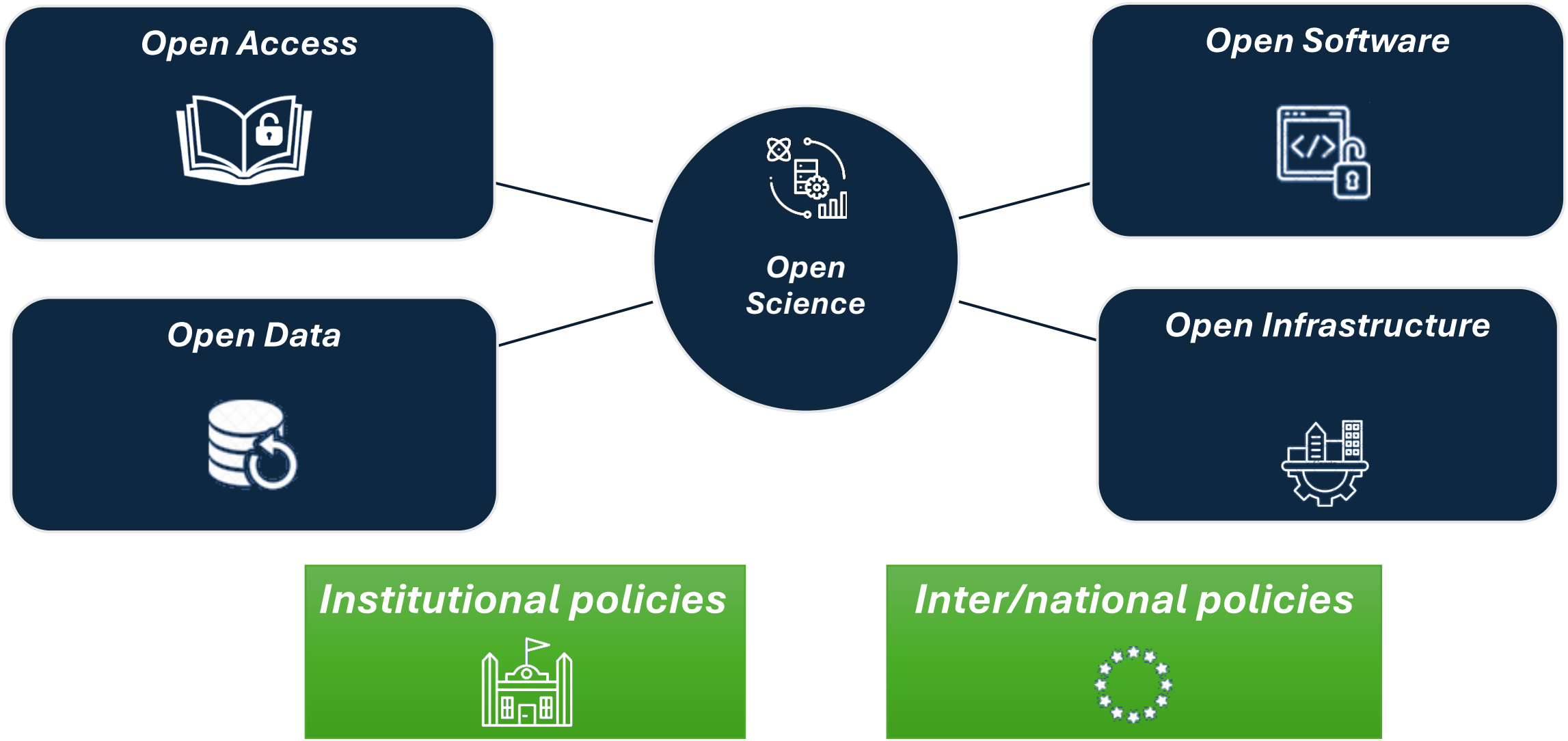
Offers a **new metric** for research assessment to remove the outdated dependence on e.g. h-index

Strengthens **technology transfer** with industry partners

On the agenda of many **governments and funding bodies**

Ultimately, the researcher should benefit from making research outputs open

Pillars of Open Science



Open Science Policies/Statements

Number of Open Science polices as

broader statements

- Can be broad OS policy or separate

Policy for each pillar

Institutional

Inter/national

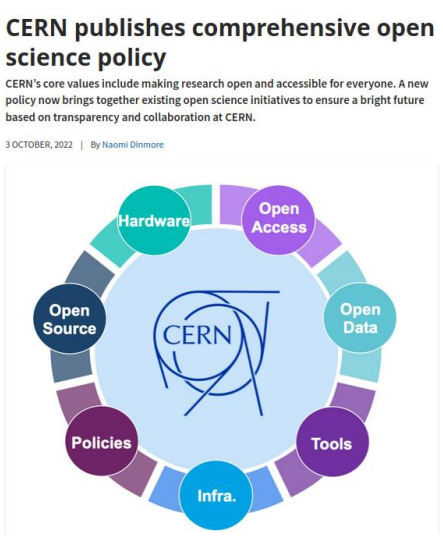
Chapter on OS included

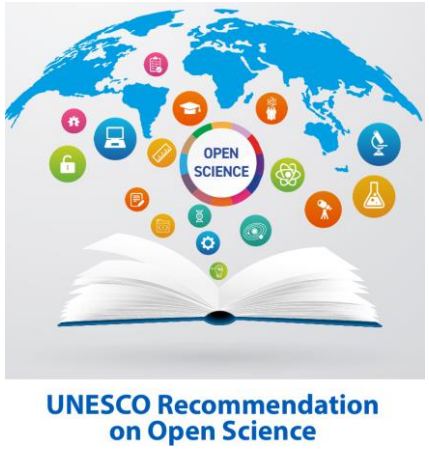
in Strategic Plans

NuPECC LRP

APEC

ECFA





Literature: Moradi & Abdi., **Open Science-related policies in Europe**, *Science and Public Policy*, (2023) https://doi.org/10.1093/scipol/scac082

F.A.I.R Principles outlines

Findable

- Centrally orchestrated storage and access of data essential to enable the data/software to be findable.
- Usage of Persistent IDentifiers (PID), Digital Object Identifiers (DOI) -> Guarantee access to Digital Research Objects.
- Generation of ‘data record’ in discipline specific repositories

Accessible

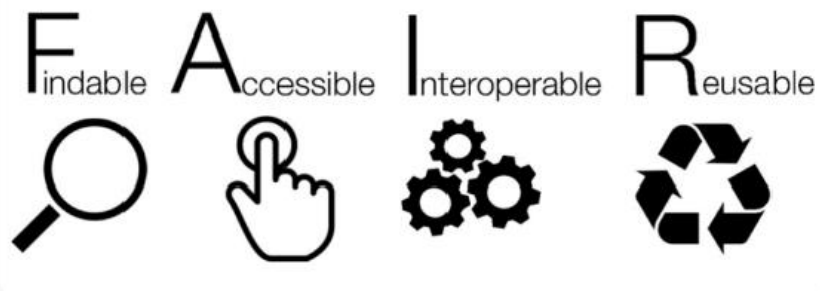
- Data and software produced/dedicated for F.A.I.R communities and publications centrally stored.
- Common & “user-friendly” interface to store and retrieve data

Interoperable

- Common metadata formats
- F.A.I.R-produced data operable with other datasets

Reusable

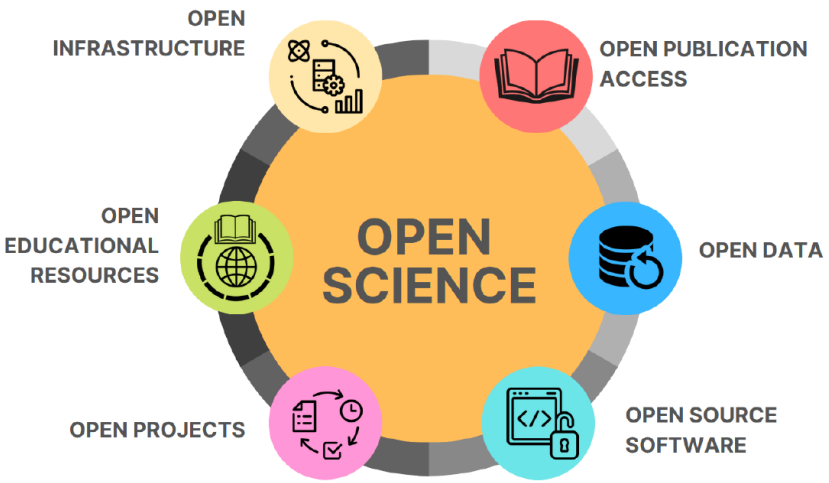
- Ensure (as reasonably as possible) data stored long term
- Metadata should be retained indefinitely



https://www.go-fair.org/fair-principles/ -> See here for more detailed info

Open Science : What we want to achieve ?

- **Open Access Publications** –> Mandatory publication of Open Access articles
- **Research Data** –> Publish research data in suitable repositories (F.A.I.R. Data)
- **Open Software** –> Make open whenever possible (F.A.I.R. Software)
- **Open Infrastructure** –> Open Projects in research and industry
- Develop an **Open Science Ecosystem** to combine everything



Considerations:

- The steps and processes to achieve this are complex… Start smaller and work up
- Do not make it too ‘general’ needs finer granularity and use-cases
- Aim to address **all researchers in the nuclear physics community** : Students, Postdocs, PI’s, Group leaders…

Foreword : Digital Objects

Data are any digital objects :

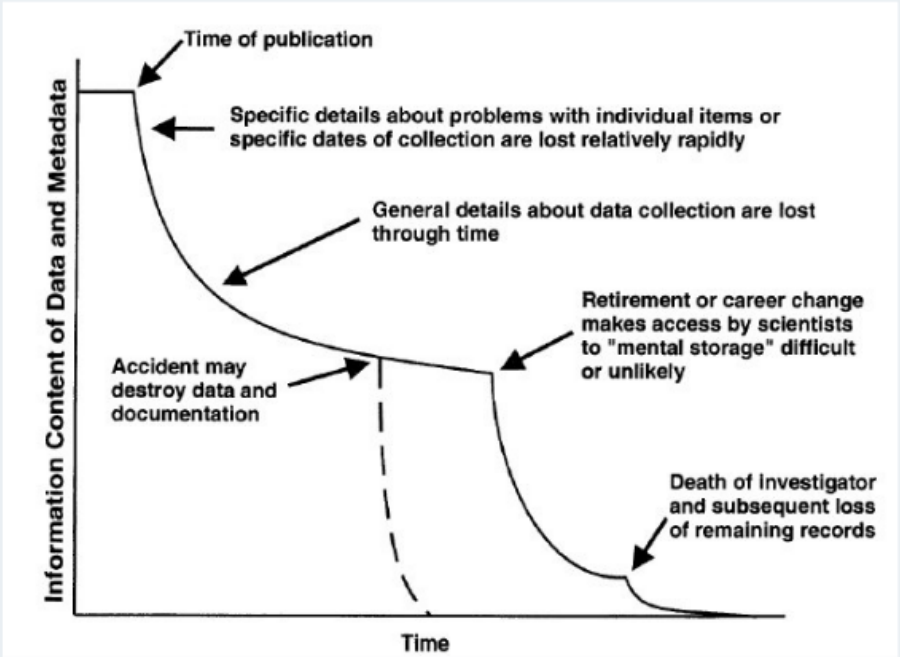
- Experimental Data-sets (raw, auxiliary-data, refined, ...)
- Simulations, Results of calculations
- Databases
- Software (sources code, Workflows, ...)
- Reports, Publications, Slide-Shows, Websites, Pictures, ...
- Data Management Plans !
- ...

Data have a Life Cycle !

Picture from Research Data Lifecycle by LMA Research Data Management Working Group

Foreword : Why should we care about Data Management ?

Data and Metadata Entropy



W. K. Michener et al., Eco. App. 7 (1997) 330-342

No data set is perfect and self-explanatory

- Too often relying on human/mental storage
- Crucial to accurately interpret results and their origin (from processing, analysis, and modeling)
- Accessibility and Reproducibility of research results
- Enhance visibility of research within and outside research domain

Long term Preservation and Management

- Defining data policies (access, sharing, preservation period, ...)
- Re-Use opportunities, Facilitate Cross Domain research
- How to chose if a data-set should be kept (Unlimited storage erea is behind us !)

Foreword : FAIR Data Practices

Findable :

- F1 Data are assigned a globally unique and persistent identifier
- F2 Data are described with rich meta data
- F3 Meta data clearly and explicitly included the identifier of the data they described
- F4 (Meta)data are registered or indexed in a searchable resource

Accessible :

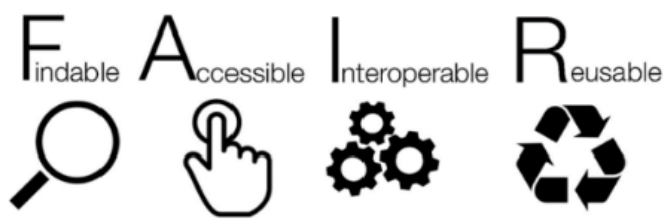
- A1 (Meta)data are retrievable by their identifier using a standardized communication protocol
- A1.1 The protocol is open, free and universally implementable
- A1.2 The protocol, where necessary, allows for an authentification & authorisation procedure
- A2 Metadata are accessible, even when the data are no longer available

Inter-operable :

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

Reusable :

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



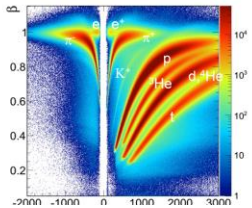
Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)

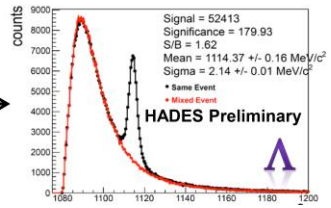
How FAIR are we/you ?

Required critical exercice of our/your level of data FAIRness

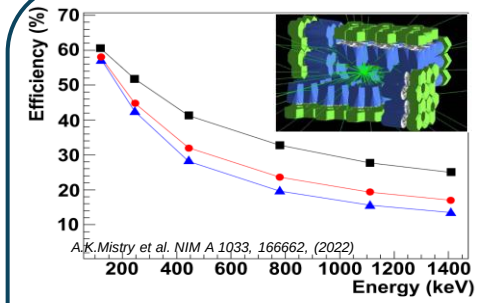
Research Data and Software : Rich and varied

Experimental Data:

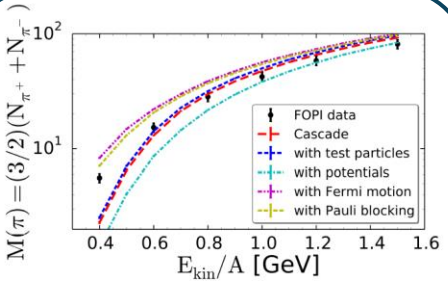




Raw *Pre-processed* *Result*



Simulation



Theory

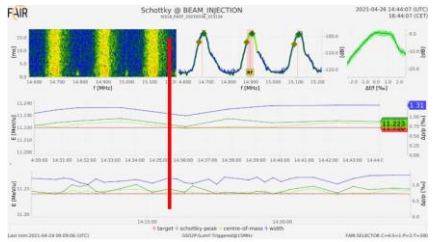
```
template <typename Modus>
void Experiment<Modus>::run() {
  const auto &mainlog = logg[LMain];
  for (event_ = 0; !is_finished(); event_++) {
    mainlog.info() << "Event " << event_;

    // Sample initial particles, start clock, some printout
    initialize_new_event();

    run_time_evolution(end_time_, {});

    if (force_decays_) {
      do_final_decays();
    }
```

Software



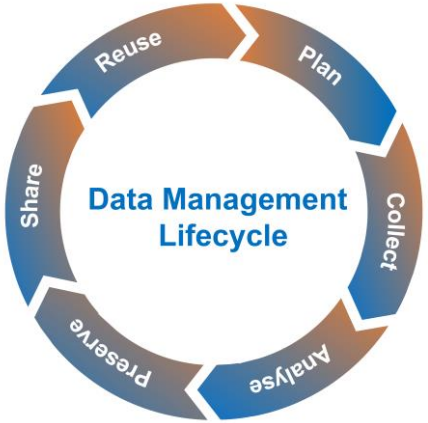
Machine Parameters

Volumetry varies widely from one community/experiment to another

Open Research Data and Management

Research Data Management encompasses all aspects of handling research data, from planning, its generation and processing to publication, long- term archiving, and eventual deletion, while adhering to the principles of good scientific practice.

One of the crucial philosophies of RDM are the F.A.I.R principles, which follow "as open as possible, as closed as necessary"

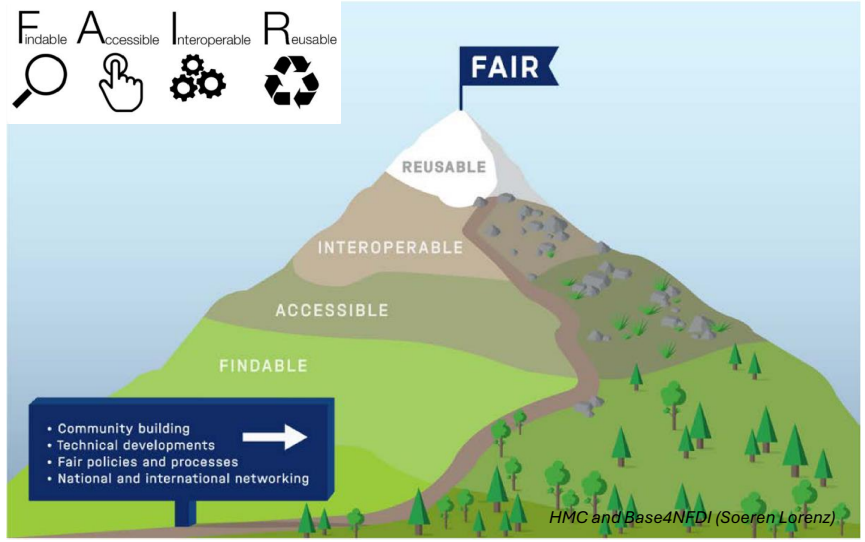


FAIR Data is not an end goal

-> continual process of improving practices and adapting research resources with technology innovations

Goals:

- to ensure good RDM practices
- promote and assist researchers in publishing data;
- to aim (as best as reasonably possible) that data is published according to the Findable Accessible Interoperable and Reusable principles;
- develop the tools and infrastructure needed to do this (repositories, Electronic logbooks ...)



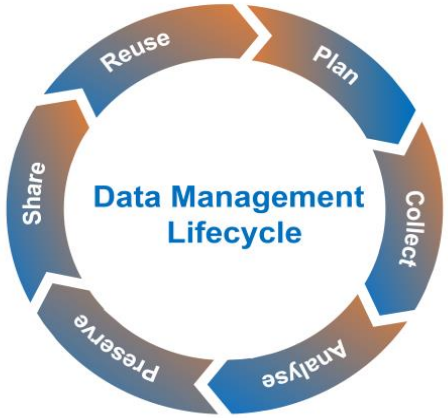
Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al*. **The FAIR Guiding Principles for scientific data management and stewardship**. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Data Management Plans

- A **Data Management Plan** is a research project document that aids in the process of ensuring that research data is handled correctly. Describes the Data lifecycle of the project
- **Living document** -> Should be filled out at the start of the project and continuously updated throughout
- Reluctance to filling out Data Management Plans! Seen as just more paperwork…
- **BUT**: Useful for researchers (present and future), needed to enable F.A.I.R data, many funding agencies now require a DMP to be prepared at the start of the project etc.

Data management plans aid:

- Communication tool for researchers
- RDM project internal management
- Future reuse and project planning
- Useful for IT/Resource Cost estimates
- Funding body requirements



Contents can include:

- General info (Project name/PI)
- Data info (Data type, size, generation method)
- Overhead (Data protection, personnel costs…)

Goal: Make it easy and encourage to prepare these

General statements about Open Software

Open Software strongly connected with Open Methods and **Open Data**(and more …)

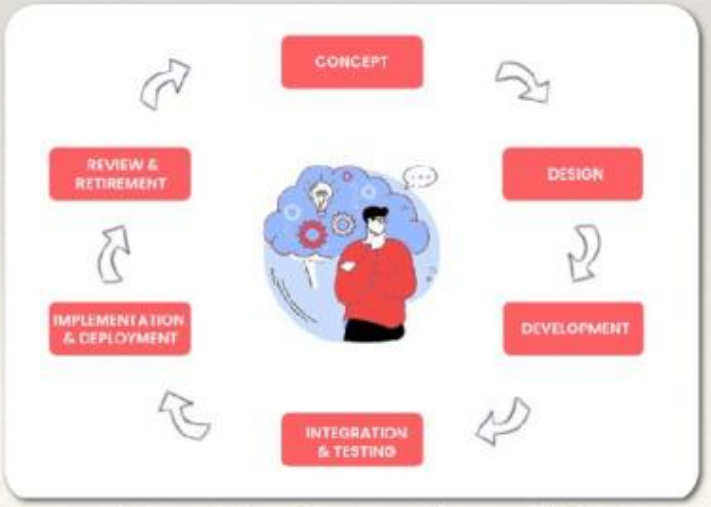
https://opensocialwork.org/research/open-science/

Methods are associated with **software**

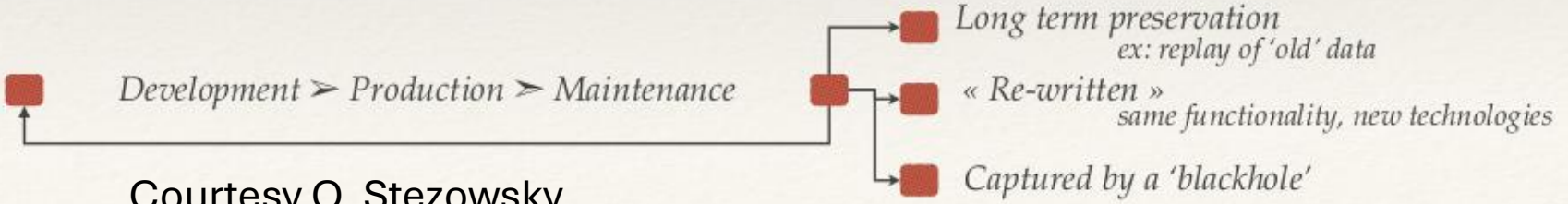
- It requires expertise
- This expertise has to be kept !

Pieces of code are **data** ! But specific :

- Written by humans (so far !)
- Need to be run
  Interpreted, compiled (built, linked) …
- Need hardware / software (OS) support
  Pb of portability, obsolescence …



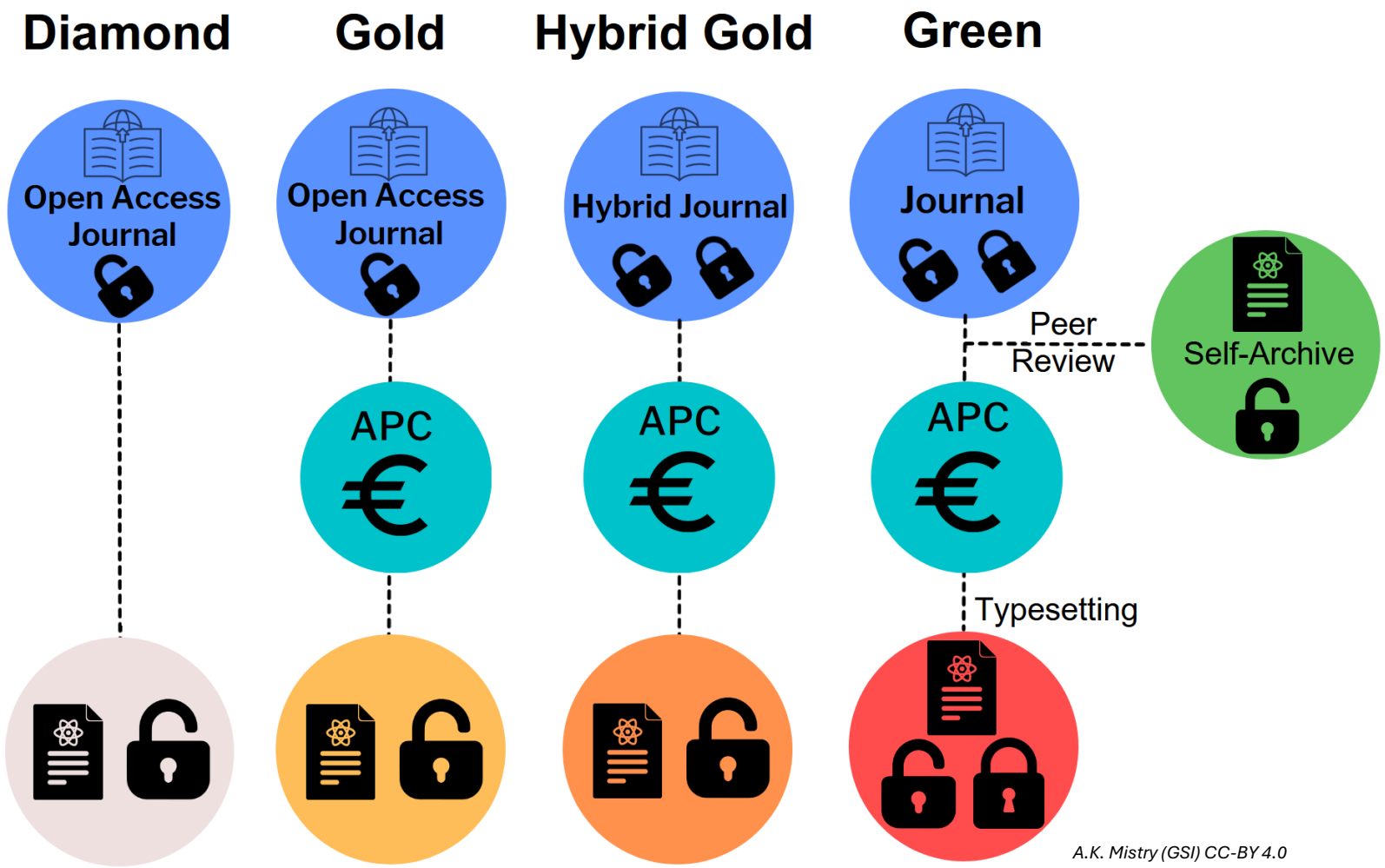
Any piece of software has a lifetime !



Courtesy O. Stezowsky

Open Access Publication Policies

- Should encourage open access publication 100% -> NuPECC LRP can strongly support this
- Already a requirement in many institutes



Infrastructures for an effective Open Science

- Data Storage (Data Lakes)
- Data Access Management (AAI)
- Data/Software Repositories and Catalog (zenodo, …)
- Software Forges (gitlab, github, …)
- Analysis Platforms
- Publication Repositories (arxiv, …)

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures

A reliable **distributed** data infrastructure capable of managing **Exabyte-scale** data sets, able to deliver data **on-demand** at l**ow latency** to all types of data processing facilities

Services operated by the ESCAPE partner institutes

Petabyte scale storage: DESY, SURF-SARA, IN2P3-CC, CERN, IFAE-PIC, LAPP, GSI and INFN (CNAF, ROMA and Napoli)

Data management and storage orchestration (Rucio)

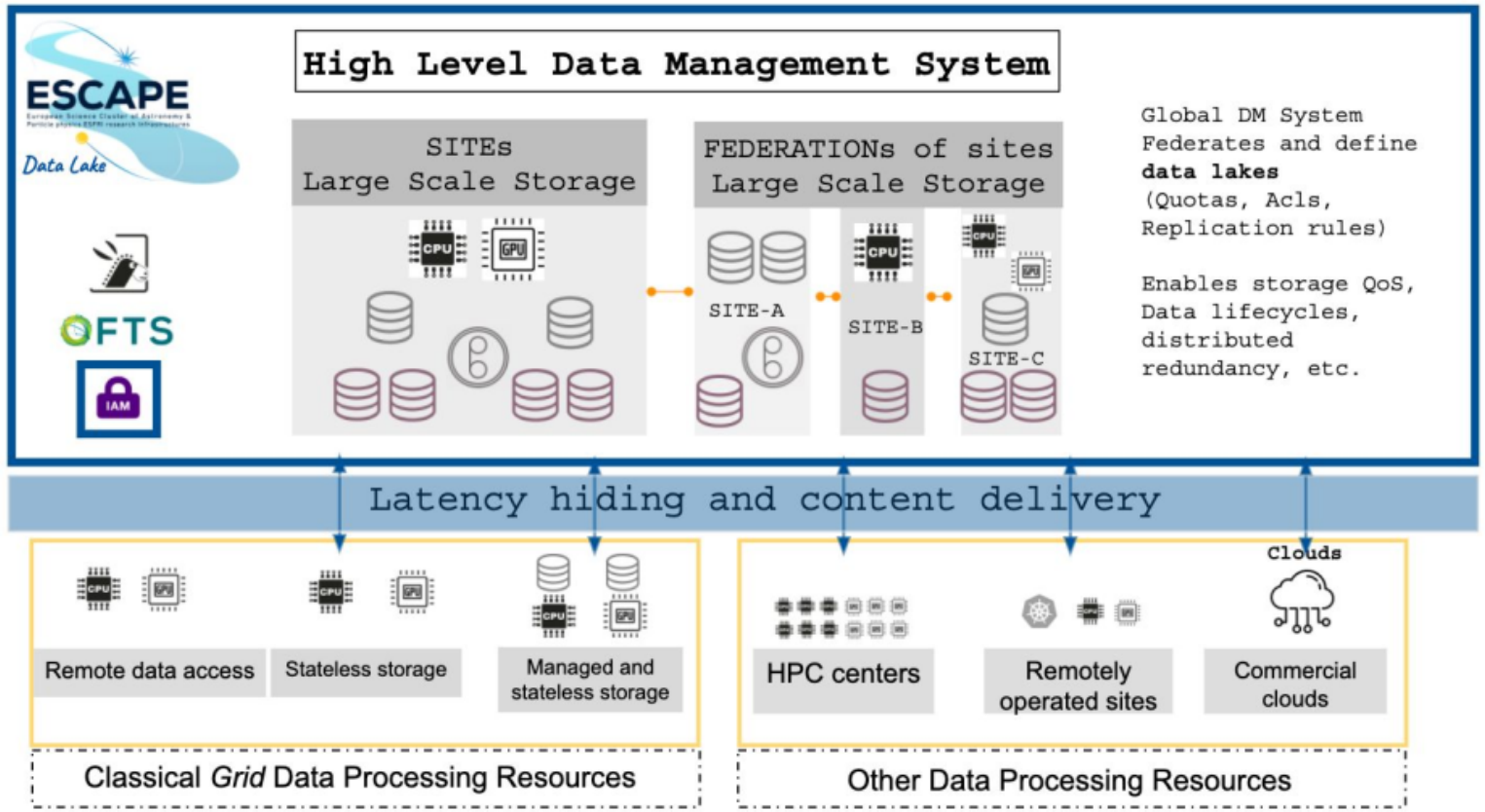
File transfer and data movement services (FTS)

Global Data Lake Information System (CRIC)

ESCAPE IAM: common Auth/Authz/IM (AAI)

Content delivery and latency hiding: XCache

Widening access with several data access protocols: http, xrootd and gridftp.



Open science : A long journey ahead of us

Opportunities

- Adoption of good practices in Data Management, Software, Publications
- Improve drastically management and visibility of data-sets
- Preservation of scientific products (data, software, analysis workflows, publications, …)
- Reproducibilty of results
- Develop new collaborations based on combined or reused data-sets.
- Enhance the scientific impact of the available and future data-sets

Required steps

- Training and promotion of good practices
- Coordination of the practices across domains
- Devellopement of tools and infrastructures (Data Lakes, Analysis Plateforms, Standardized Metadata, …)
- Recognition of these activities

Require strong involvement of all stakeholders (Researchers, RI, IT departments …)

Open Science require a cultural change in our daily work

- From the researcher point of view
- From the infrastructure point of view

Hands On Session

- Instruction to install the environment :
https://gitlab.in2p3.fr/eurolabs-os-school/teaching/project-zero

Prerequisites

Micromamba environment

All dependencies for the project could be installed through micromamba using the following command:

```
"${SHELL}" <(curl -L micro.mamba.pm/install.sh)
curl "https://gitlab.in2p3.fr/lpc-dev/misc-examples/micromamba/-/raw/main/root_py11.yaml?ref_type=hea
micromamba env create -f root_py11.yaml
```

To activate the environment use:

```
micromamba activate root_py11
```

- Data will be provided by usb keys by Jérémie and Adrien