DPHEP & ICFA Data lifecycle panel:

Recommendations for data preservation and open science in particle physics

RHIC roundtable- June 18, 2026

Cristinel Diaconu (CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France)

Kati Lassila-Perini (Helsinki Institute of Physics, Finland)

Ulrich Schwickerath (CERN)

We have preserved data!!!

High-energy physics (HEP) facilities
Data types

High-energy physics (HEP) data

DESY (Hamburg, Germany)		
JADE (PETRA)	1979–1986	1 TB
H1, ZEUS (HERA)	1992–2007	700 TB

CERN (Geneva, Switzerland)		
ALEPH, DELPHI, L3, OPAL (LEP)	1989–2000	400 TB
ALICE, ATLAS, CMS, LHCb (LHC)	2010–2041	O(1 EB)

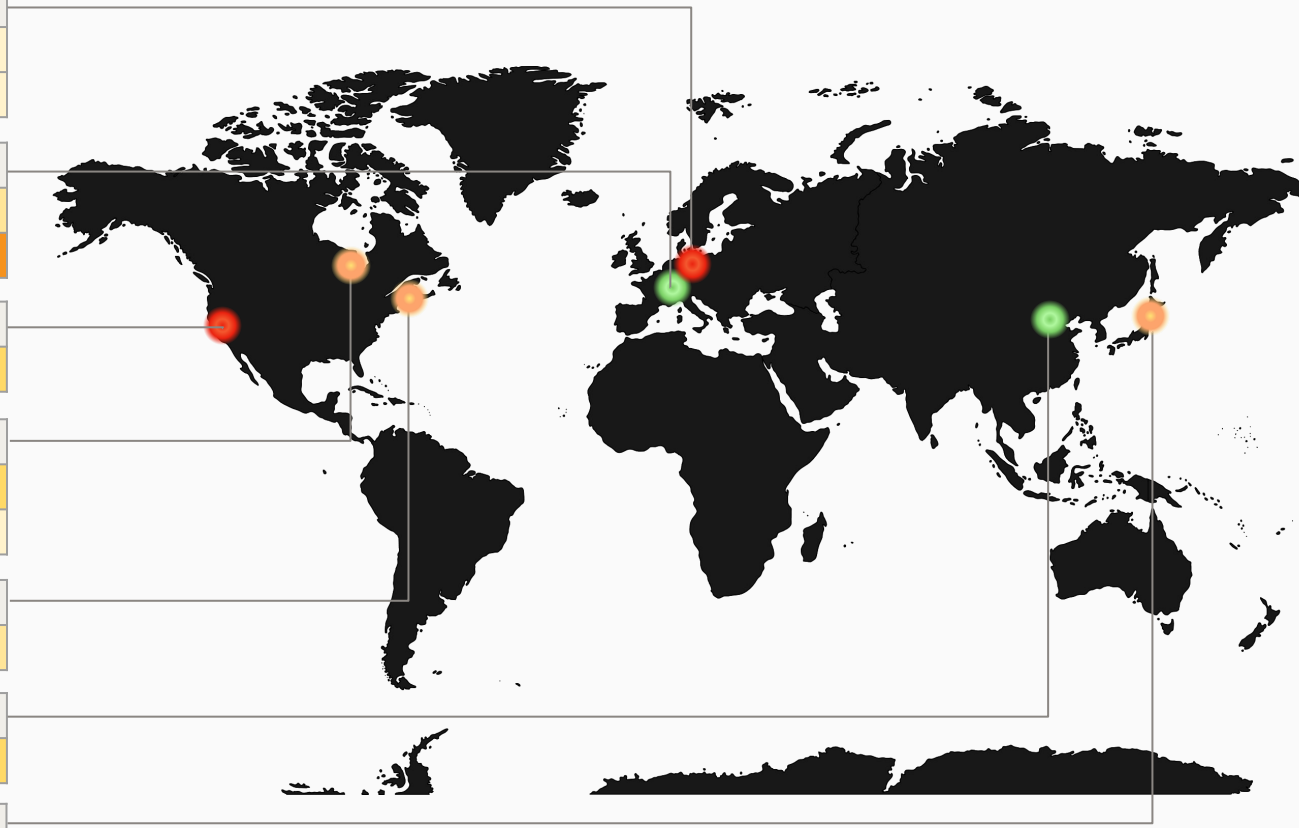
SLAC (Stanford, CA, USA)		
BABAR (PEP II)	1999–2008	2 PB

FERMILAB (Batavia, IL, USA)		
CDF, H1 (TEVATRON)	1983–2011	17.5 PB
MINERVA (ν-beam)	2010–2019	10 TB

BNL (Brookhaven, NY, USA)		
PHENIX (RHIC)	2000–2016	25 PB

IHEP (Beijing, China)		
BES III (BEPCII)	2009–2030	6 PB

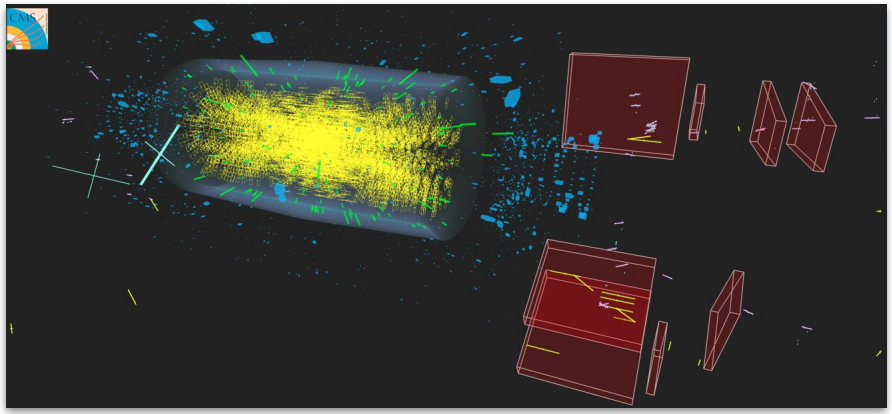
KEK (Tsukuba, Japan)		
Belle I (KEKB)	1999-2010	4 PB



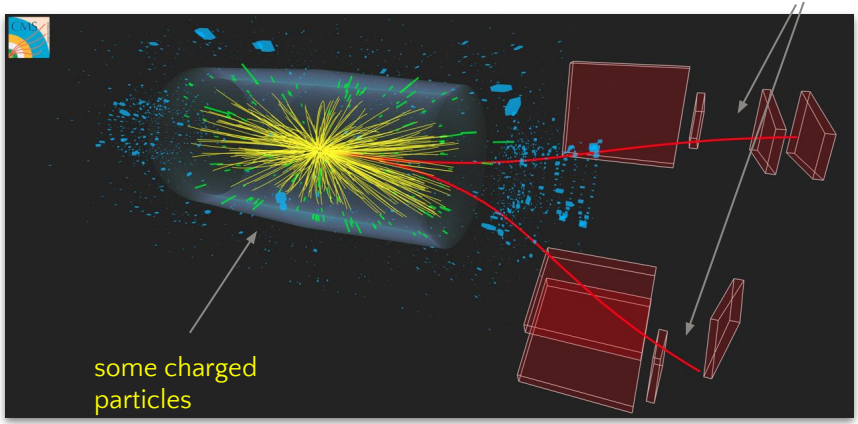
Facility (location)		
Experiment (Collider)	data taking	preserved data

Data types: “Event data”?

Single collision event as raw data…



… as reconstructed data

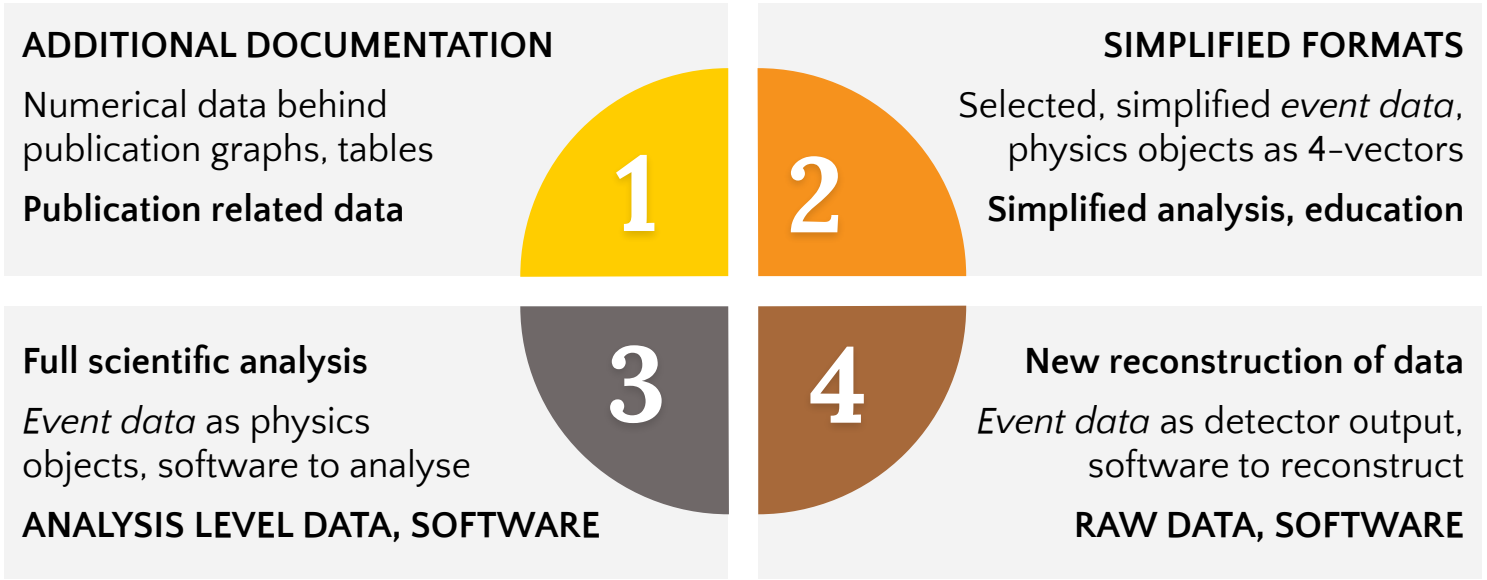


To set the scale (LHC, currently operating at CERN):

Event size:	raw → reconstructed: O(1MB) → O(100kB) → O(10kB) → O(1kB) (various formats)
Collision rate:	40MHz → O(1k) events selected and stored (unbiased selection = “something happened”)
Beam time:	O(100d/y) = O(10M s/y)
	→ O(10 G evts/y) → O(10 PB/y)

Image credit: CMS collaboration, T Mc Cauley

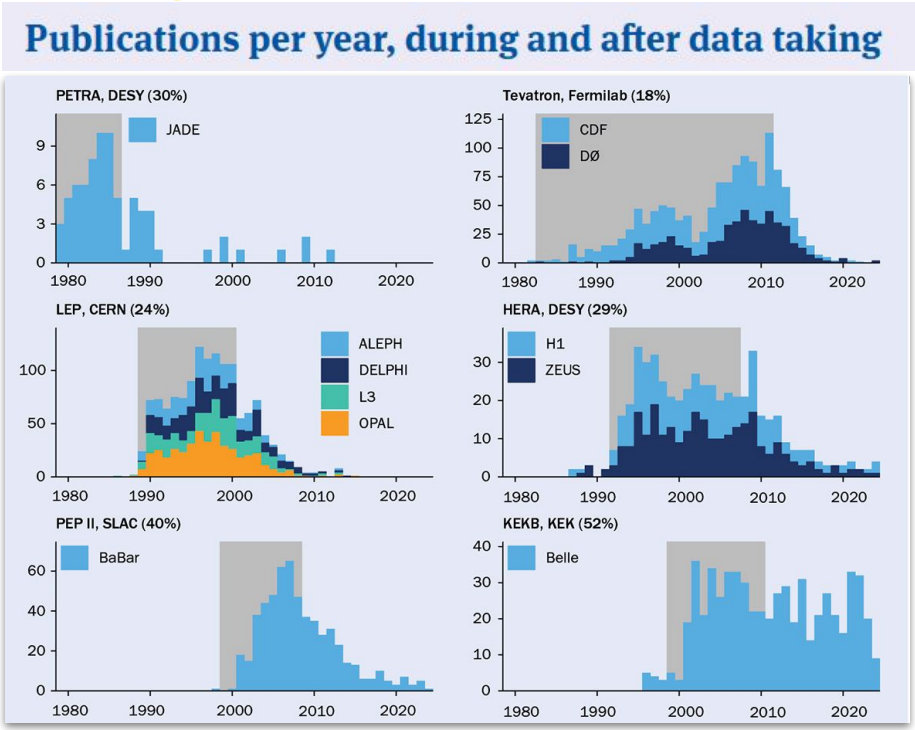
HEP data preservation levels



Preserved data in use

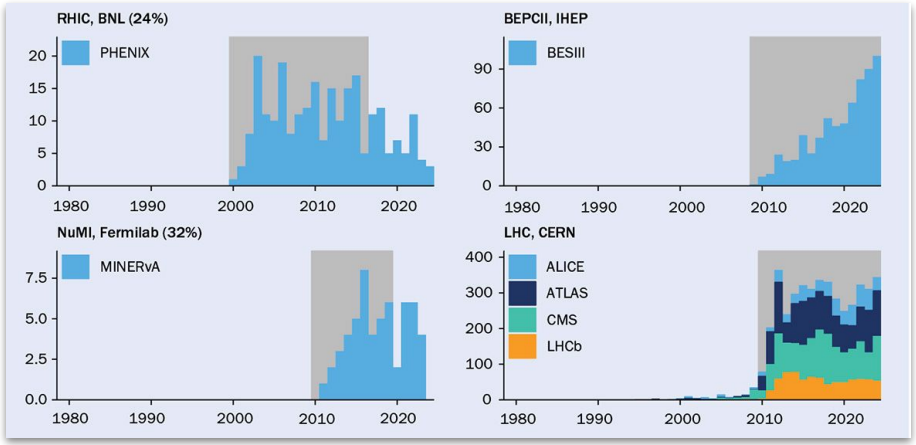
Usage within the original collaboration
Usage for new purposes

Scientific output using preserved data



Within the collaborations. Note:

- long time span of data taking (▭)
- % indicates N of publications after data taking

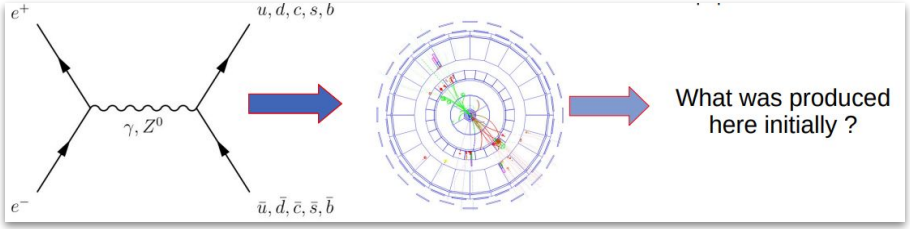


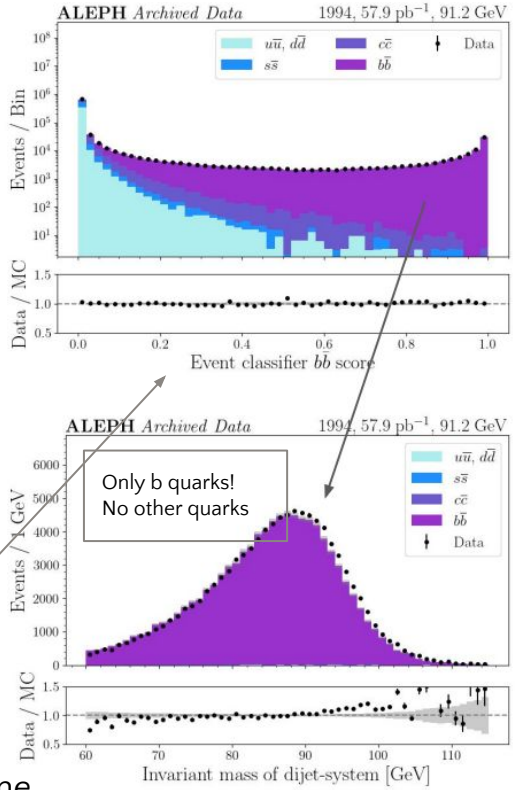
Source: Hidden treasures, CERN Courier, image credit: D De Biasio/ M Rayner

New studies with preserved data

Often with **researchers who were involved in the original collaboration**

Example: Particle identification
Using old collider data for future collider studies





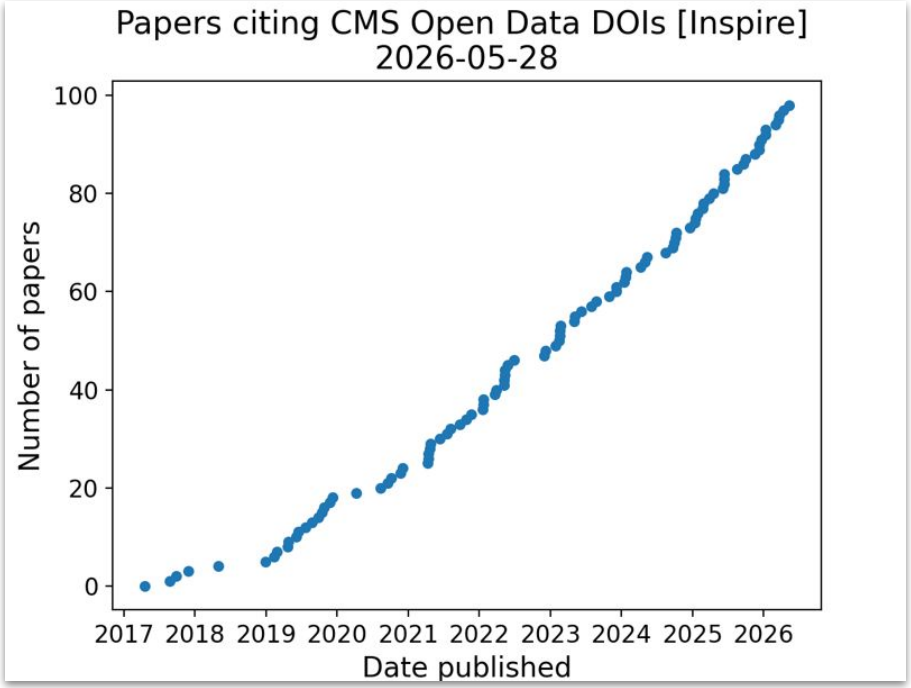
e^+e^- 1994 data: Aleph (LEP)

→ use new advanced analysis techniques
→ way better than anything that was possible at the time.

→ testbed for future colliders (FCC-e^+e^- 2040s)

Source:
Advanced jet flavour tagging with archived ALEPH data
5th DPHEP workshop, April 2026

New studies with preserved open data

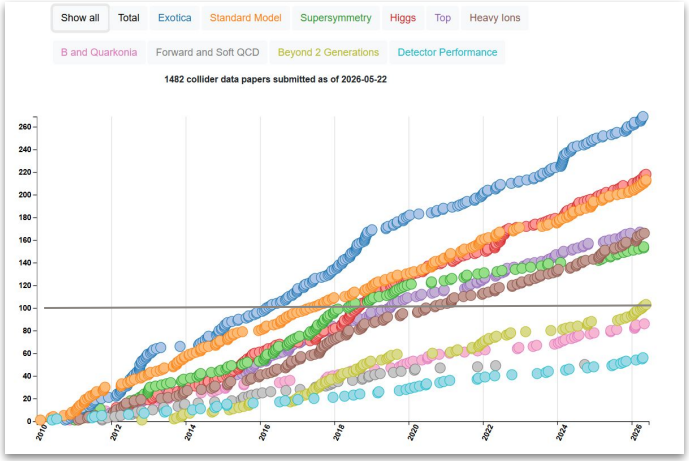


Source: https://doi.org/10.5281/zenodo.15078670

Also **by researchers who are not involved in the original collaboration**

Example: CMS at the LHC

Compare to publications by the CMS experiment:



Source: CMS publicatione timeline – Credit: CMS collaboration

Open science: Policies, commitments, outcomes

CERN Open Science Policy

The CERN Open Science Policy covers all elements of the Open Science relevant to CERN. This includes, in particular open access to research publications, data, software and hardware, as well as research integrity, infrastructure, education and outreach activities supporting or enabling open science practices.

> Supported by long term financial investments from its Member and Associate Member States, with significant contributions also from non-Member States, CERN is committed to the advancement of science and the wide dissemination of knowledge by embracing and promoting practices making scientific research more open, collaborative, and responsive to societal changes. ... CERN accordingly recognizes the holistic practice of open science as one of its guiding principles.

Last revision: Oct 2022

CERN Open Access Policy

The CERN Open Access Policy defines the principles and processes through which CERN authors can publish their peer-reviewed articles Open Access. A dedicated website also provides authors with additional resources to find the easiest route to comply with the policy.

> CERN authors are required to publish all of their peer-reviewed primary research articles open access (by default under a Creative Commons attribution license, i.e. CC-BY-4.0)

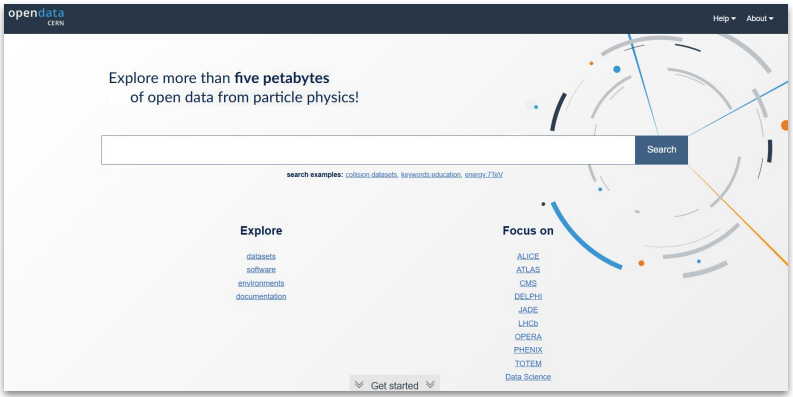
Last revision: May 2021

CERN LHC Open Data Policy

This policy relates to the data collected by the LHC experiments for the main physics programme of the LHC — high-energy proton-proton and heavy-ion collision data. The foreseen use cases of the Open Data include reinterpretation and reanalysis of physics results, education and outreach, data analysis for technical and algorithmic developments, and new physics research.

> Making data available responsibly, at different levels of abstraction and at different points in time, allows the maximum realisation of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community.

Last revision: Nov 2020



Example: CERN

Try yourself!

Launch a notebook:

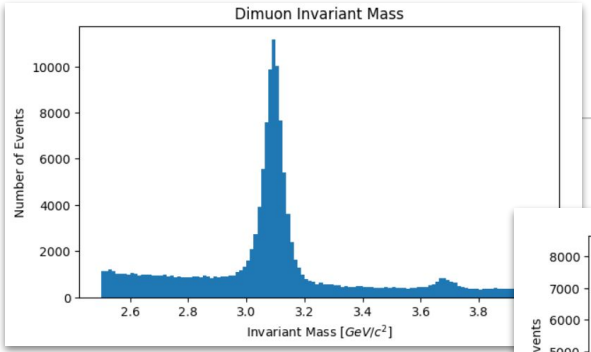
launch binder

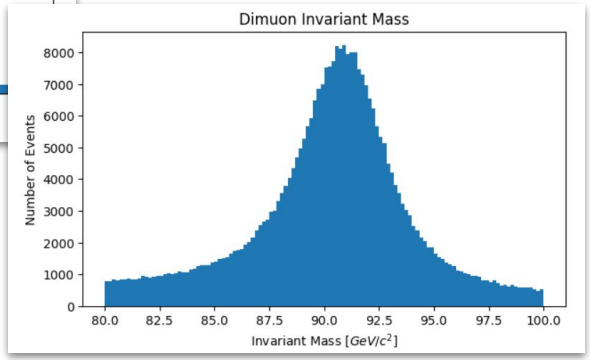
Find some Nobel prize winning particles:

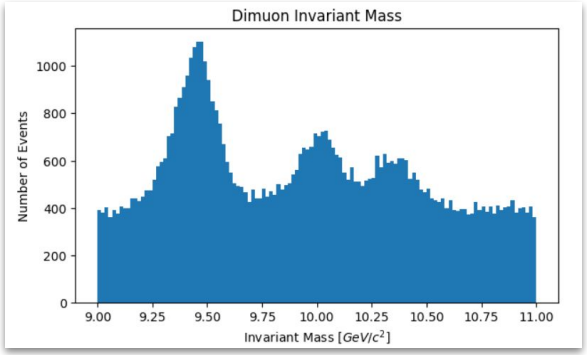
- J/ψ, Z

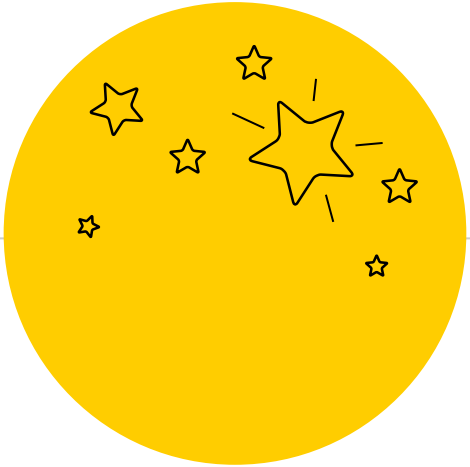
Not a Nobel prize, but a great discovery and a good story:

- Upsilon









Data preservation and open science?

We are seeing great progress.

What makes it happen?

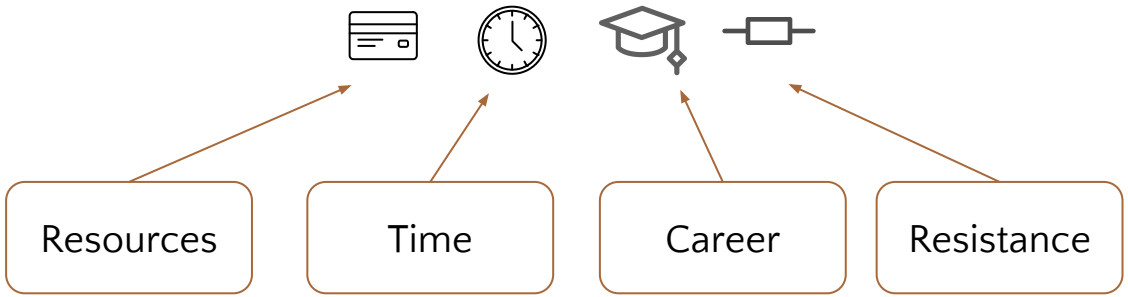
A structured approach or ...

Magic?

Data preservation and open science:

...people with good will and necessary skills in positions in which they can do what they believe should be done despite of challenges!

Challenges related to:



How to overcome the challenges?

Domain-specific actors
How to encourage integrating data and knowledge preservation to everyday work

Domain-specific actors

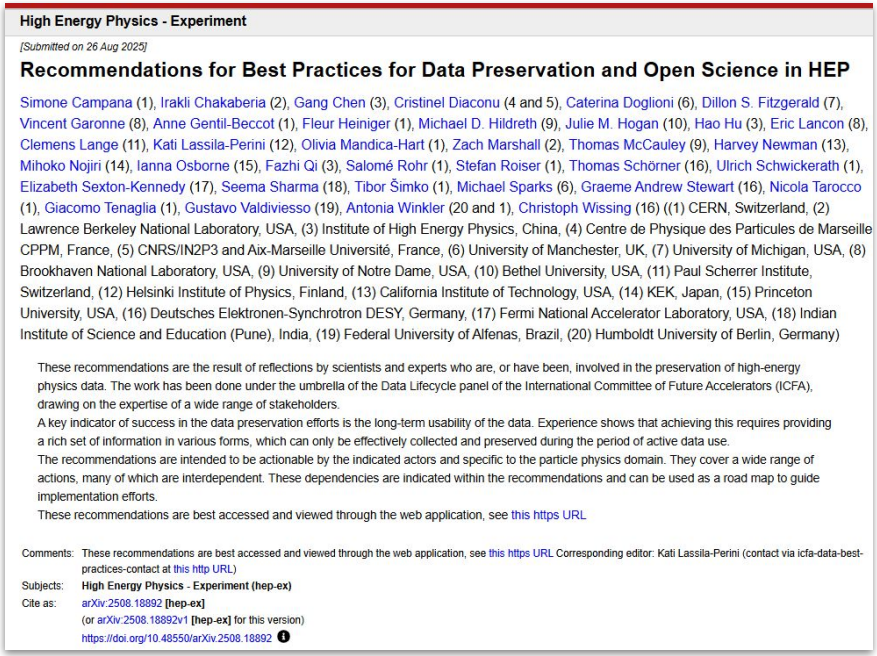
Data Preservation in High-Energy Physics (DPHEP)

- Collaboration of those active and interested in DP
- Sharing knowledge (workshops), status reports

Data Lifecycle panel of the International Committee for Future Accelerators (ICFA)

- Working group to enhance global coordination with focus on Open Science and FAIR practices
- As a first major task, formulating community-driven recommendations

Recommendations



High Energy Physics - Experiment

[Submitted on 26 Aug 2025]

Recommendations for Best Practices for Data Preservation and Open Science in HEP

Simone Campana (1), Irakli Chakaberia (2), Gang Chen (3), Cristinel Diaconu (4 and 5), Caterina Doglioni (6), Dillon S. Fitzgerald (7), Vincent Garonne (8), Anne Gentil-Beccot (1), Fleur Heiniger (1), Michael D. Hildreth (9), Julie M. Hogan (10), Hao Hu (3), Eric Lancon (8), Clemens Lange (11), Kati Lassila-Perini (12), Olivia Mandica-Hart (1), Zach Marshall (2), Thomas McCauley (9), Harvey Newman (13), Mihoko Nojiri (14), Ianna Osborne (15), Fazhi Qi (3), Salomé Rohr (1), Stefan Roiser (1), Thomas Schörner (16), Ulrich Schwickerath (1), Elizabeth Sexton-Kennedy (17), Seema Sharma (18), Tibor Šimko (1), Michael Sparks (6), Graeme Andrew Stewart (16), Nicola Tarocco (1), Giacomo Tenaglia (1), Gustavo Valdiviesso (19), Antonia Winkler (20 and 1), Christoph Wissing (16) ((1) CERN, Switzerland, (2) Lawrence Berkeley National Laboratory, USA, (3) Institute of High Energy Physics, China, (4) Centre de Physique des Particules de Marseille CPPM, France, (5) CNRS/IN2P3 and Aix-Marseille Université, France, (6) University of Manchester, UK, (7) University of Michigan, USA, (8) Brookhaven National Laboratory, USA, (9) University of Notre Dame, USA, (10) Bethel University, USA, (11) Paul Scherrer Institute, Switzerland, (12) Helsinki Institute of Physics, Finland, (13) California Institute of Technology, USA, (14) KEK, Japan, (15) Princeton University, USA, (16) Deutsches Elektronen-Synchrotron DESY, Germany, (17) Fermi National Accelerator Laboratory, USA, (18) Indian Institute of Science and Education (Pune), India, (19) Federal University of Alfenas, Brazil, (20) Humboldt University of Berlin, Germany)

These recommendations are the result of reflections by scientists and experts who are, or have been, involved in the preservation of high-energy physics data. The work has been done under the umbrella of the Data Lifecycle panel of the International Committee of Future Accelerators (ICFA), drawing on the expertise of a wide range of stakeholders.

A key indicator of success in the data preservation efforts is the long-term usability of the data. Experience shows that achieving this requires providing a rich set of information in various forms, which can only be effectively collected and preserved during the period of active data use.

The recommendations are intended to be actionable by the indicated actors and specific to the particle physics domain. They cover a wide range of actions, many of which are interdependent. These dependencies are indicated within the recommendations and can be used as a road map to guide implementation efforts.

These recommendations are best accessed and viewed through the web application, see this https URL

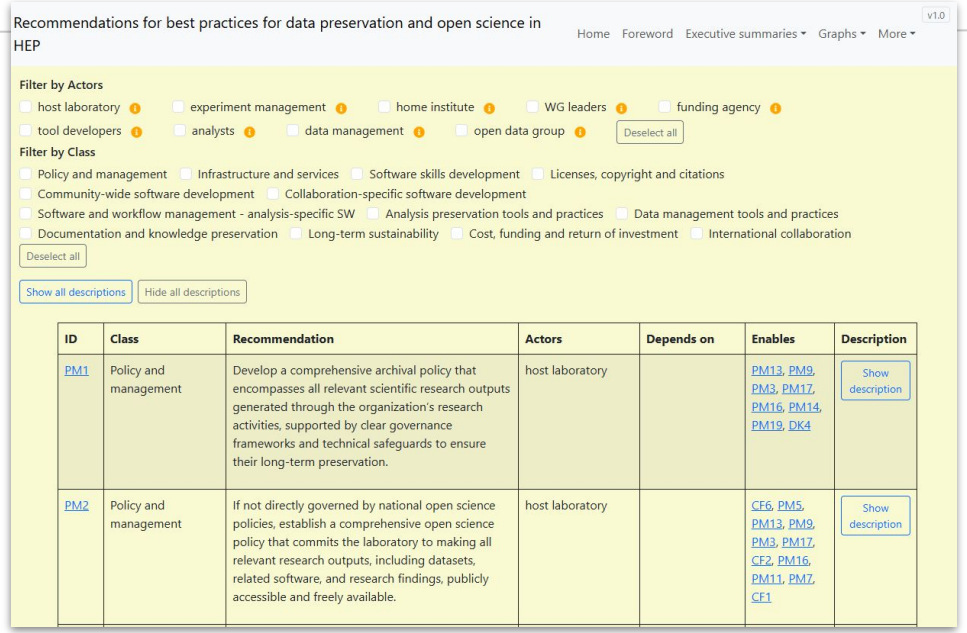
Comments: These recommendations are best accessed and viewed through the web application, see this https URL Corresponding editor: Kati Lassila-Perini (contact via icfa-data-best-practices-contact at this http URL)

Subjects: High Energy Physics - Experiment (hep-ex)

Cite as: arXiv:2508.18892 [hep-ex]

(or arXiv:2508.18892v1 [hep-ex] for this version)

https://doi.org/10.48550/arXiv.2508.18892



Recommendations for best practices for data preservation and open science in HEP

Home Foreword Executive summaries Graphs More v1.0

Filter by Actors

host laboratory | experiment management | home institute | WG leaders | funding agency

tool developers | analysts | data management | open data group | Deselect all

Filter by Class

Policy and management | Infrastructure and services | Software skills development | Licenses, copyright and citations

Community-wide software development | Collaboration-specific software development

Software and workflow management - analysis-specific SW | Analysis preservation tools and practices | Data management tools and practices

Documentation and knowledge preservation | Long-term sustainability | Cost, funding and return of investment | International collaboration

Deselect all

Show all descriptions | Hide all descriptions

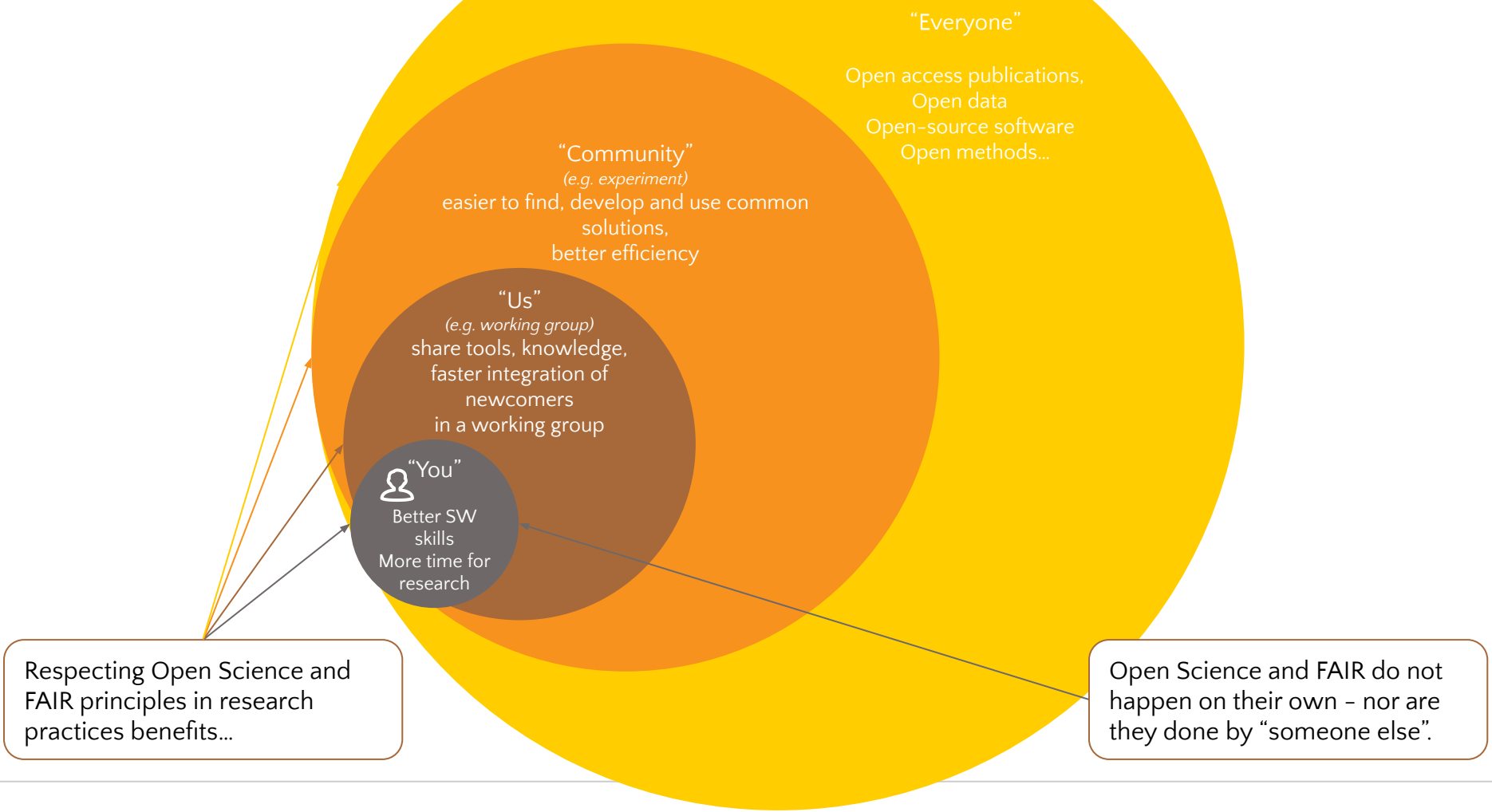
ID	Class	Recommendation	Actors	Depends on	Enables	Description
PM1	Policy and management	Develop a comprehensive archival policy that encompasses all relevant scientific research outputs generated through the organization's research activities, supported by clear governance frameworks and technical safeguards to ensure their long-term preservation.	host laboratory		PM13, PM9, PM3, PM17, PM16, PM14, PM19, DK4	Show description
PM2	Policy and management	If not directly governed by national open science policies, establish a comprehensive open science policy that commits the laboratory to making all relevant research outputs, including datasets, related software, and research findings, publicly accessible and freely available.	host laboratory		CF6, PM5, PM13, PM9, PM3, PM17, CF2, PM16, PM11, PM7, CF1	Show description

https://icfa-data-best-practices.app.cern.ch/
https://arxiv.org/abs/2508.18892
ICFA statement

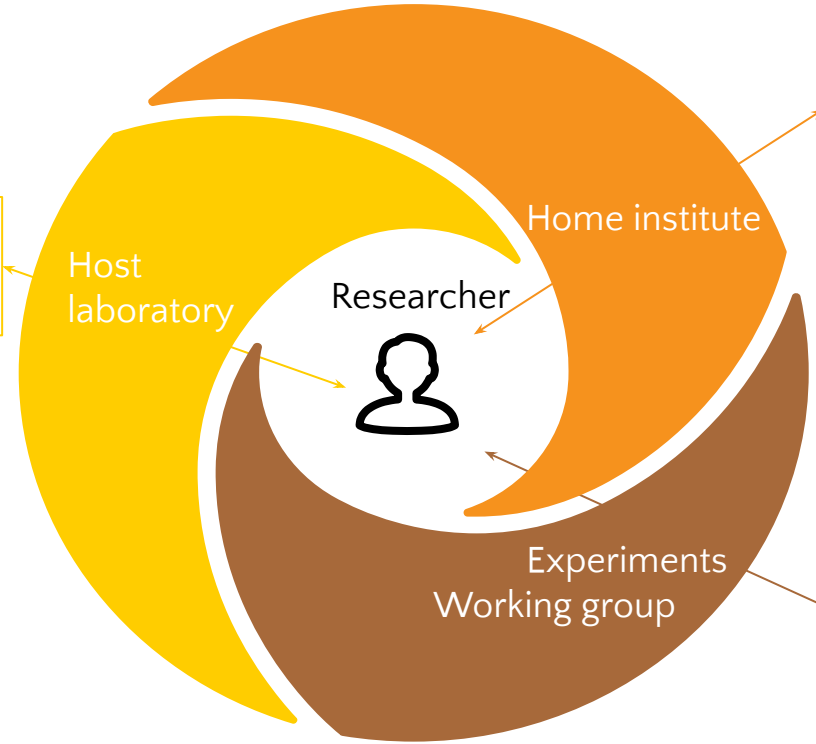
Why this matters?

Benefits of respecting FAIR principles in research?

FAIR: Findable, Accessible, Interoperable, and Reusable digital assets (data, software, workflows, ...)


"Everyone"
Open access publications,
Open data
Open-source software
Open methods...
"Community"
(e.g. experiment)
easier to find, develop and use common solutions,
better efficiency
"Us"
(e.g. working group)
share tools, knowledge,
faster integration of newcomers
in a working group
"You"
Better SW skills
More time for research
Respecting Open Science and FAIR principles in research practices benefits...
Open Science and FAIR do not happen on their own – nor are they done by "someone else".

Many stakeholders



Task definition, share of time on them
Resources (salary, travel, personal tools)
Education, training
Research/data management guidelines

Infrastructure
Resources
Open science guidelines

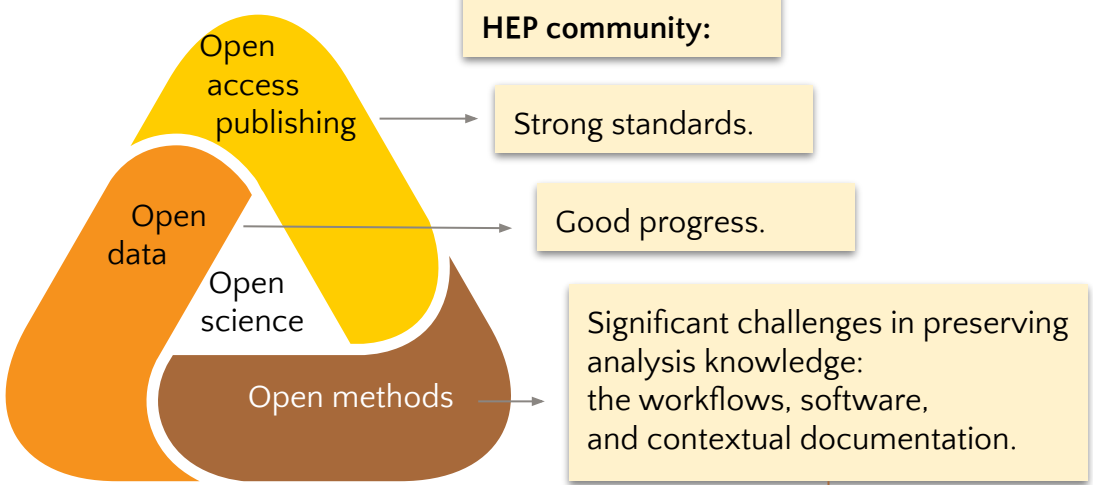
Responsibilities
Resources (computing, software)
Practices ("how do we work")
Policies

Recommendations

Design principles
Key takeaways

The goal and the current situation in HEP

Key indicator of success: **long-term scientific usability of experimental data.**



Design principles

Recommendations for best practices for data preservation and open science in HEP

Home Foreword Executive summaries Graphs More

Filter by Actors

host laboratory experiment management home institute WG leaders funding agency tool developers analysts

data management open data group Deselect all

Filter by Class

Policy and management Infrastructure and services Software skills development Licenses, copyright and citations Community-wide software development

Collaboration-specific software development Software and workflow management - analysis-specific SW Analysis preservation tools and practices Data management tools and practices

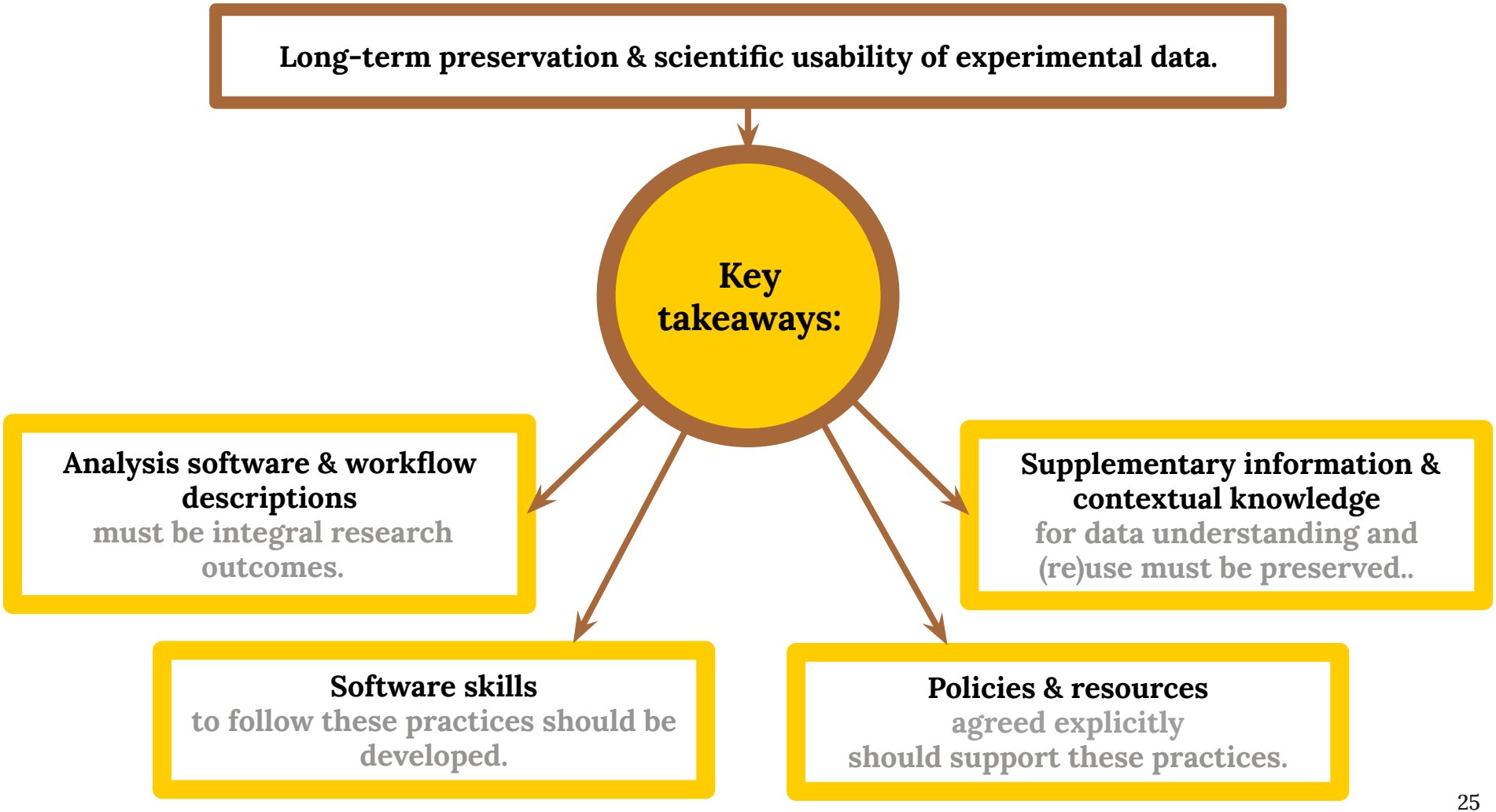
Documentation and knowledge preservation Long-term sustainability Cost, funding and return of investment International collaboration Deselect all

Show all descriptions Hide all descriptions

ID	Class	Recommendation	Actors	Depends on	Enables	Description
SW2	Software and workflow management - analysis-specific SW	From the early stages of an analysis, store your code in findable and accessible repositories within the version control infrastructure designated by the experiment.	analysts	SW1, IR1	SW11, SW9, SW8	Hide description
		Motivation Common analysis code repositories within an experiment make analysis code findable and accessible to all members of the experiment. This facilitates the use of common tools, sharing knowledge, and transparent analysis review.				
SW3	Software and workflow management - analysis-specific SW	Use code versioning tools (e.g. Git), and commit changes frequently to the common repository.	analysts	SK3, IR1	SW4, AP7, SW11	Hide description
		Motivation Code versioning ensures a clear history of changes, and enables easy rollback to previous versions. Frequent commits help catch issues early, make your progress visible, and reduce conflicts by integrating small, manageable updates regularly.				

For the best readability, select an actor role

Short recommendation, further description/motivation text available


Long-term preservation & scientific usability of experimental data.
Key takeaways:
Analysis software & workflow descriptions
must be integral research outcomes.
Supplementary information & contextual knowledge
for data understanding and (re)use must be preserved..
Software skills
to follow these practices should be developed.
Policies & resources
agreed explicitly should support these practices.

Next?

Reach different actor groups
Evaluate how different actors implement these recommendations
Report and share solutions

Follow-up

Within the community and with neighbouring fields

Seminars, presentations

Start with the decision-level actor groups that make possible actions by others

- **Host laboratories**
- **Experiment managements**

Assessment

Status options:	The assessment team should…
Applied	provide links to the relevant information (document/services/instructions)
Partially applied	provide links to the relevant information (document/services/instructions) indicate what is missing
Planned	indicate the timeline
Not yet considered	indicate reasons for not being yet considered. indicate the timeline to be considered.
Not applicable	indicate the reasons why

Use this information: **to incentivize dialogue between experiments, and the host laboratory to improve and update the recommendations if needed.**

Assessment - outcome?

Raise awareness:
regular assessments rounds at ~2y intervals
sync with management changes
start with an introduction session
highlight benefits

Share solutions:
focus on the overall picture
why certain actions not taken
identify common challenges
highlight successful actions
no emphasis on specific shortcomings

Monitor progress:
compare to previous rounds
encourage change
keep the impact positive

Summary: From community experience to actionable recommendations

- Data preservation and open science efforts are gaining traction and bringing benefits in the high-energy physics community, but
  - sustainability challenges persist
  - actions are not always systematically taken during all data lifecycle
- Actionable best-practice recommendations:
  - **Concrete, Specific & Relevant:** building on community expertise, co-created by the HEP community.
  - Stepwise roadmap to ensure long-term scientific reusability of HEP data.
- The assessment process:
  - **Raise awareness**: reach beyond data preservation aficionados.
  - **Share solutions**: highlight good practices and working examples.
  - **Incentivize dialogue**: communicate training, tool and staffing gaps.

Thank you!

Questions?
Discussion?

Presentation template from SlidesCarnival
Icons from flaticon – authors: freepik, joalfa
Word cloud from WordArt.com