Open and FAIR – data, workflows, science



ICFA Seminar, Toyama, Japan - November 4, 2026

Kati Lassila-Perini
ICFA Data Lifecycle panel
Helsinki Institute of Physics - Finland

FAIR?

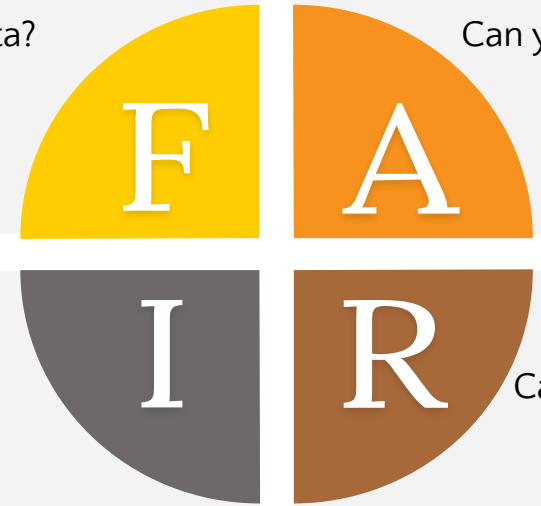
https://www.gofair.foundation/fair-principles

FINDABLE

Do you know where to look for data?
Can you find what you need?

ACCESSIBLE

Can you download or access otherwise?
Openly, or with authentication if required?



Are they in some common format?
Do you have the tools to analyze them?

INTEROPERABLE

Do you know how to use them?
Can you do new research with them?

REUSABLE

Data + software + workflows → science

Data:

collected data

simulated data

beam conditions

luminosity information

data quality information

data selection

calibrations

corrections

scale factors

physics object definitions

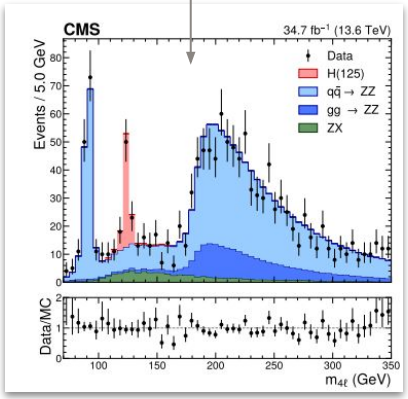
...

Software:

community-wide sw

experiment-specific sw

analysis-specific sw



http://dx.doi.org/10.1007/JHEP05(2025)079

Workflows

(= steps in data analysis):

not documented

in a readme file

in scripts

in workflow languages

(by AI agents)

some of these

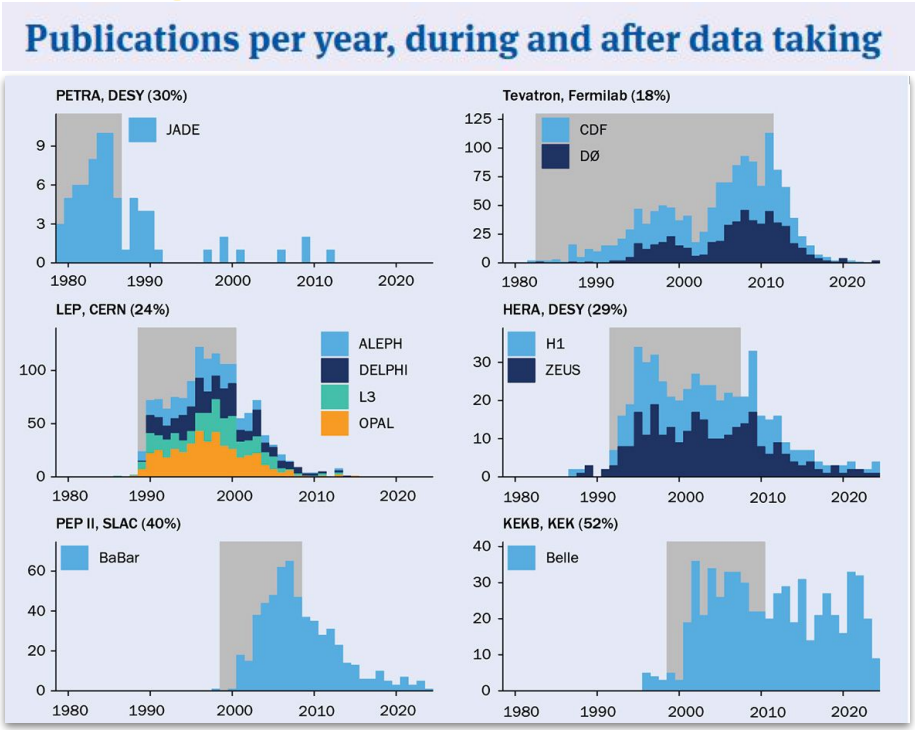
Computing:

environment

dependencies

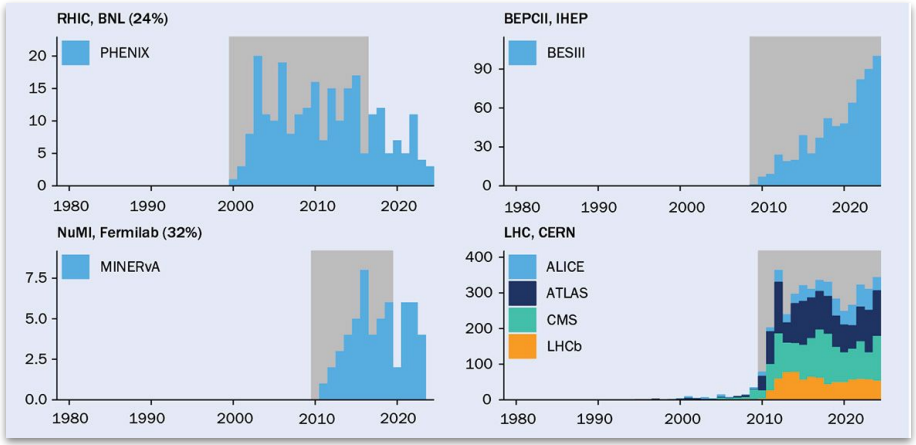
OS / SW versions

Scientific output using preserved data



Within the collaborations. Note:

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

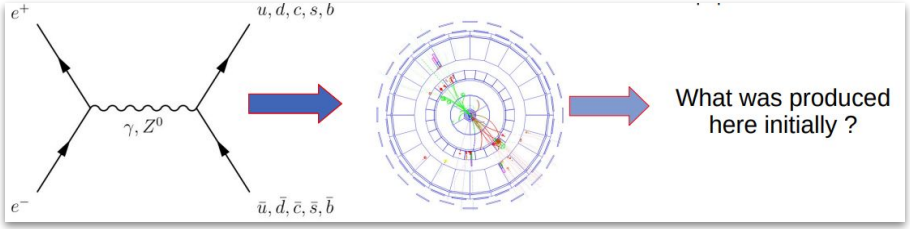


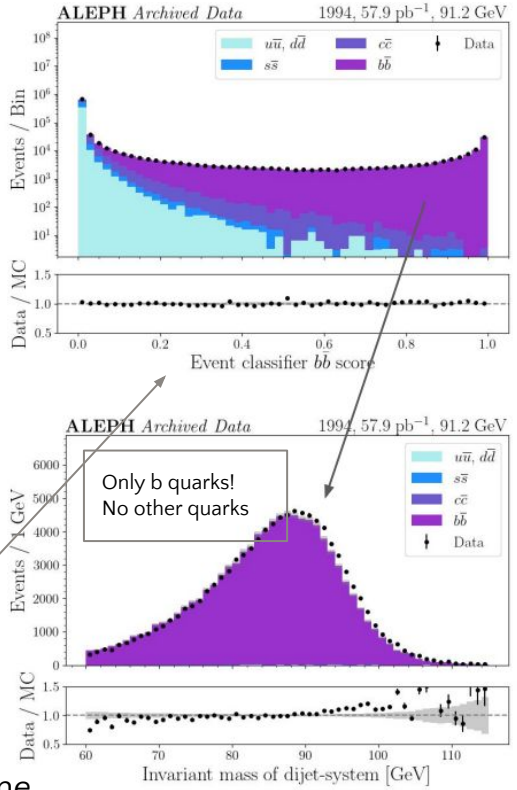
Source: Hidden treasures, CERN Courier, image credit: D De Biasio/ M Rayner

New studies with preserved data

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

Example: Particle identification
Using old LEP data for FCC studies





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

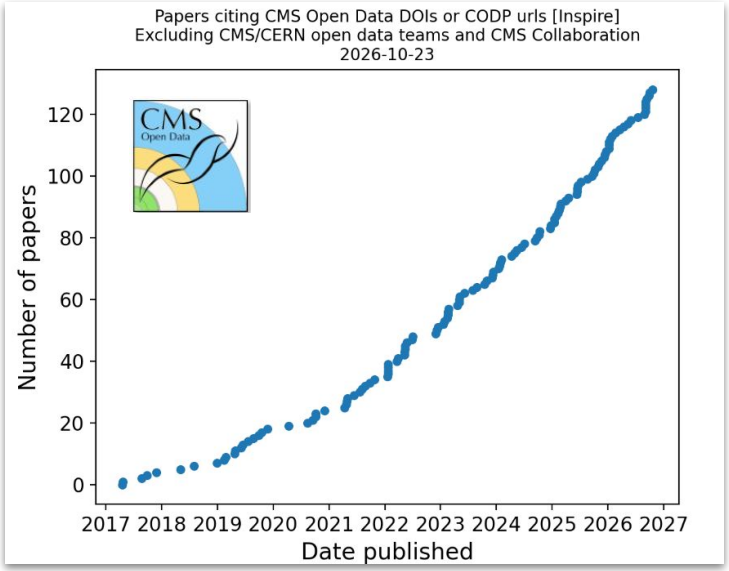
→ use new advanced analysis techniques
→ way better than anything that was possible at the time.

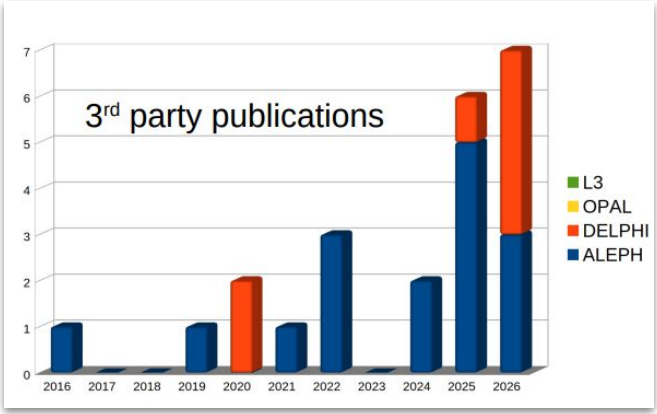
→ testbed for future colliders (FCC-e^+e^- 2040s)

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

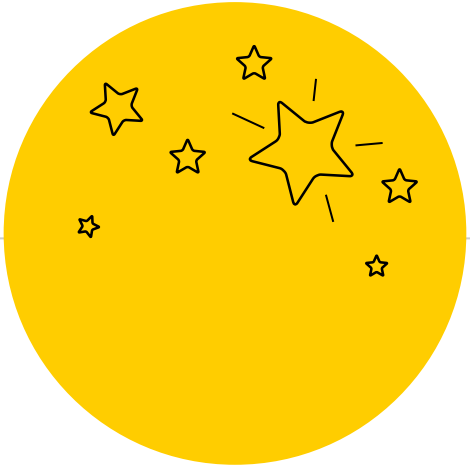
FAIR → Open Science:
New studies with preserved open data

By researchers not in the original collaborations





CMS publication plot: https://doi.org/10.5281/zenodo.23599198
LEP publications, private communication Ulrich Schwickerath, updated version of
https://indico.cern.ch/event/1529317/contributions/6946554/attachments/3232116/5762245/greetings-from-the-past.pdf



FAIR and open science?

We are seeing great progress.

What makes it happen?

A structured approach or ...

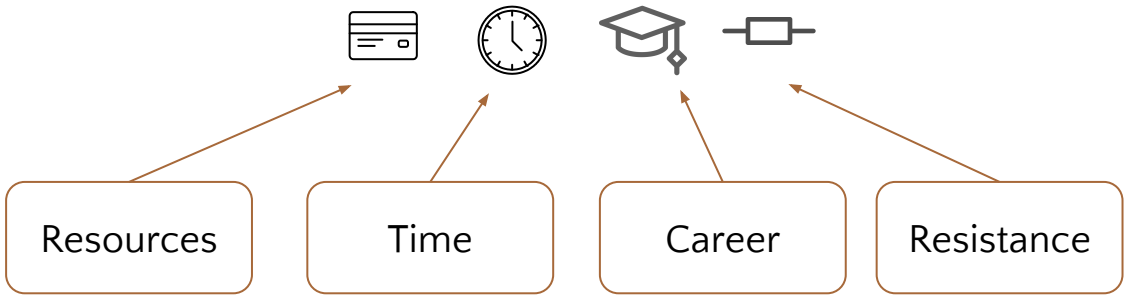


Magic?

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 challenges!

Challenges related to:



How to overcome the challenges? Domain-specific actors

Data Preservation in High-Energy Physics (DPHEP)

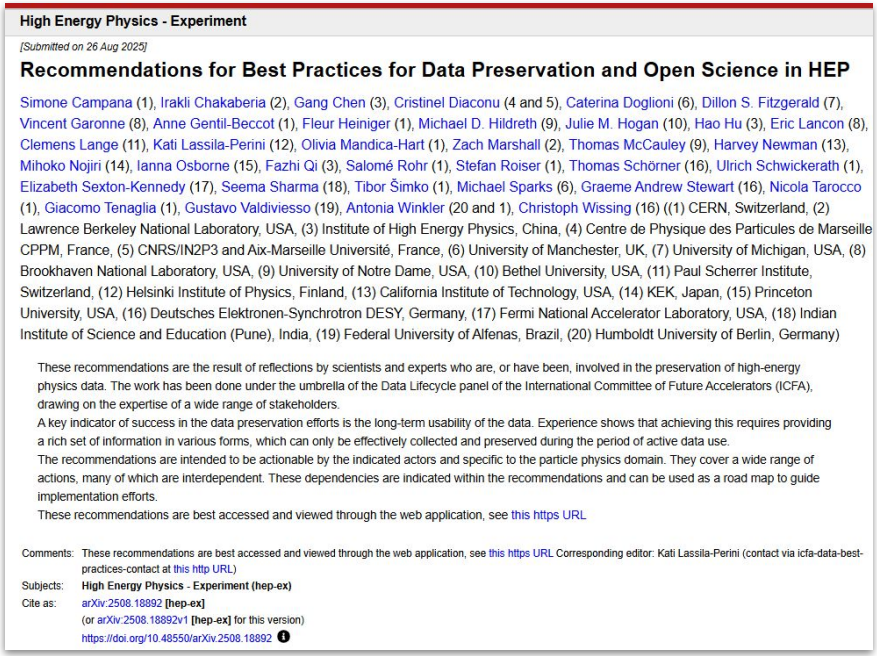
- Collaboration of those active and interested in DP
- Sharing knowledge (workshops), status reports

ICFA Data Lifecycle panel

- A working group with a focus on Open Science and FAIR practices
- As a first major task, formulate community-driven recommendations
  - Panel + scientists and experts with hands-on expertise on curating and using preserved data: **bottom-up**
  - How to integrate FAIR practices into everyday work

Outcome

Concrete, specific, actionable recommendations relevant to the indicated actor group



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

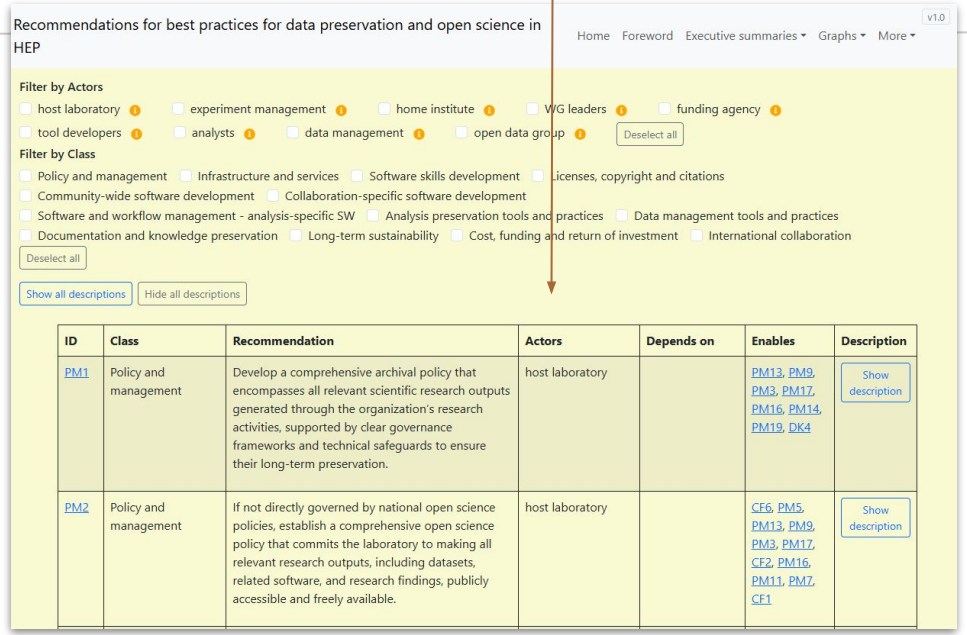
Comments: These recommendations are best accessed and viewed through the web application, see this https URL Corresponding editor: Kati Lassila-Perini (contact via icfa-data-best-practices-contact at this http URL)

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



Key points

- Preserve data
- Curate and preserve what makes them usable
  - Think beyond the lifetime of the experiment
- Consider analysis code a research outcome to be preserved
  - Make internal analysis reproducibility an expectation
  - Allow time for software management skills
- Release data for common benefit and to test their usability
  - Indicate how they can be used through licences
- Recognize the need of specific efforts and funding
  - Value the work invested in FAIR practices
- Agree on and supply adequate resources.

What next?



The panel is conducting a domain-wide evaluation.

Assessment teams from

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

to self-assess how the recommendations are implemented in their case.



Analyze the input, with the focus on the overall picture

- why certain actions were not taken
- identify common challenges
- highlight successful actions

Share solutions: a report and a thematic DPHEP workshop 2027 (tbc).



Future?

Things move fast in the AI domain...

FAIR practices are even more relevant in the context of AI evolution as the importance of curated data and knowledge is increasing.

Importance of curated data and knowledge is increasing

- Many ongoing AI activities in the HEP domain, also beyond well-established ML solutions
  - LLM-assisted information retrieval
  - context-specific coding assistance
  - agentic AI tasks.
- Well-curated digital assets (data, software, workflows, documentation) will be crucial for their success
  - data availability and machine readability
  - human readability for understanding and verification of AI-assisted outcomes.
- The curation work needs dedicated resources!

Example: Impact of curated data on information retrieval

- Many ongoing projects to ease our work through LLM-assisted information retrieval from heterogeneous knowledge bases.
- Often reported: AI-assisted response lacks quality due to
  - bad documentation
  - or lacking metadata

Inserting correct metadata at the time of information input is cheap.

If decided, implemented, mandated (as outlined in the recommendations).

Validating AI-generated outputs is expensive.

Importance of documentation and FAIR practices!

Impact of AI-assisted coding

- Software is our main tool for all steps of the data lifecycle
  - AI-assisted coding changes how we work.
  - This is a wide topic and reflections here are restricted to long-term data reusability and are the author's own views.
- Benefits for generating code are obvious, but
  - human workload is going from writing code to its validation
  - analysis knowledge preservation and skill evolution is affected, if the human author does not take full "ownership" of produced code.

Validation bottleneck

- For long-term use of preserved data, the first-hand human expertise is not available.
- How to validate AI-assisted outcomes?

A simple example: analysis workflow by AI agents - AI agents will find a luminosity value, or apply some corrections e.g. to measured E values.

It is up to the humans to judge if they were correct - this is only feasible if that information is stored in a well-defined location.

Well-curated information serves as the source of truth.



Tools lifecycle and choice of solutions

- Building tools is becoming easy, but they need to be owned, operated and maintained
  - important for tools for open data, often individual initiatives and not part of mainstream code development.
- Awareness is needed for the tools' lifecycle:
  - where will they be deployed?
  - who will maintain them after the original developers leave?
  - what is their cost of use?
- For many tasks, the best long-term solution remains
  - a direct query on a well-curated information base (instead of chatbots)
  - a deterministic automated workflow (instead of probabilistic AI agents).

Impact of AI-assisted coding on data long-term usability

- ICFA recommendations put a strong emphasis on preserving the analysis code to preserve the analysis knowledge
  - not having it preserved is a well-known obstacle to data reuse
- How does “vibe coding” affect the analysis-specific code?
  - as producing code becomes effortless, will we observe a proliferation of narrow-scoped and verbose analysis code (instead of common tools)?
    - despite claims, agentic AI workflows are neither easily readable nor truly reproducible
    - the code review and validation for analysis code are not as structured as they are for common tools
- What do we consider acceptable as a community?
  - Today's researchers with a solid knowledge of software are gaining efficiency with LLMs.
  - But: how do future scientists learn to validate the code if they never learnt to code?
  - Do we really want to pay AI companies for “Software as a Service”? Or “analysis as a service”?

Research ethics and scientific integrity

- We are in a new situation: the dominance of AI companies and their intrusion into education and academic work.
- While embracing the new opportunities, we should not ignore that these companies
  - have huge power and are a threat to democracy
  - have demonstrated illegal practices and are a risk to data privacy
  - have significant environmental impacts
  - are there to make money, and their strategy is to make users dependent on their products.
- Is this a moment for a community-wide reflection on our choices?
- Echoing concerns raised by the previous speakers:
  - “Who gets access to the new tools, and how researchers learn their craft?” – Borut Paul Kerševan
  - “What scientific capabilities, human expertise, and training ecosystem must HEP preserve and develop as AI capabilities rapidly change?” – Koji Hashimoto

Reflection is starting

From the CERN AI Strategy:



Everything we do is guided by the following principles:

AI as a core capability	We place AI on the same strategic level as computing, data, and engineering.
Innovation and long-term vision	We balance immediate optimisation and sustained exploratory AI research to enable future scientific breakthroughs.
Responsibility and trust	We deploy AI systems that are safe, scientifically reliable, explainable where required, and aligned with CERN's values, legal obligations, and neutrality.
Openness and collaboration	We advance Open Science while collaborating with Member States, Europe, and industry.
Selective development	We adopt mature external solutions where suitable and focus internal effort where CERN's mission and infrastructure provide unique value.
Sustainability by design	We are mindful of energy use, lifecycle costs, and long-term maintainability in all AI development and deployment decisions.

Source:
CERN AI Strategy - March 2026

In the community, e.g.:

https://indico.cern.ch/event/1739582/



Responsible AI-assisted coding and common training aspects

Wednesday 21 Oct 2026, 16:00 → 17:30 Europe/Zurich

42/3-032 (CERN)

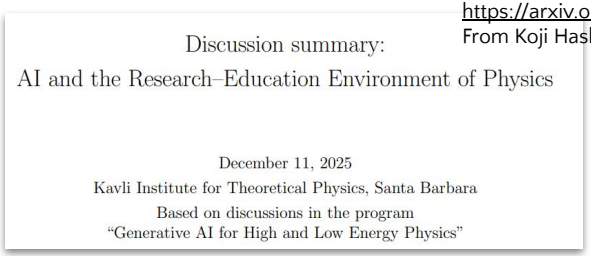
Description Contact

HSF (Caterina Doglioni. Claire Antel, Michel Hernandez Villanueva, Nicole Skidmore)

Google doc for live discussions

Supporting parties:
EVERSE + RSQKit
EuCAIF Fair & Sustainable WG
IRIS-HEP
IML
SMASH
COMCHA

https://arxiv.org/pdf/2605.03185
From Koji Hashimoto



Discussion summary:
AI and the Research–Education Environment of Physics

December 11, 2025
Kavli Institute for Theoretical Physics, Santa Barbara
Based on discussions in the program
"Generative AI for High and Low Energy Physics"

For education use of open data, see also
SCIENCE COMMUNICATION OUTREACH & OPEN SCIENCE
by Kate Shaw

Impact on Open Science

- Open data are **meant for** research-level use, for “others” to do science – e.g. in the CERN Open Data policy:
  - “The LHC experiments will release calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies.”
  - “The information provided will be sufficient to allow high-quality analysis of the data [...]”
- Data are provided with “Terms of use”
  - Datasets are made available under terms of a license waiver, other content under their applicable license conditions.
  - Are these Terms of use enough when “others” are powered by AI agents?
- So far, we have only seen responsible use of HEP open data, but will this last?

CERN Open Data Terms of Use

Documentation About

The CERN Open Data Portal is offered by CERN as part of its mission to publish the results of its experimental and theoretical work.

Use of the CERN Open Data Portal service denotes agreement with the following terms:

1. The CERN Open Data Portal is provided free of charge.
2. All metadata and datasets in the CERN Open Data Portal are made available under the terms of the CC0 waiver.
3. All other content is made available under its applicable license conditions as specified on the relevant web pages. For example, software packages are usually licensed under GNU General Public License (GPL). Download and use of such other content from the CERN Open Data Portal does not, unless expressly stated in the applicable license conditions, transfer any intellectual property. Bulk downloading of email addresses taken from CERN Open Data Portal is not allowed.
4. All content is provided “as is” and the user shall hold CERN and the content providers free and harmless in connection with its use of such content.
5. CERN reserves the right, without notice, at its sole discretion and without liability, to restrict or remove user access where it considers that use of the CERN Open Data Portal interferes with its operations or violates these Terms of Use or applicable laws.
6. These terms of Use are subject to change by CERN at any time and without notice, other than through posting the updated Terms of Use on the CERN Open Data Portal website.

If you have any questions or comments with respect to the CERN Open Data Portal, or if you are unsure whether your intended use is in line with these Terms of Use, or if you seek permission for a use that does not fall within these Terms of Use, please contact opendata-support@cern.ch.

Conclusion and outlook

- Open science – comprising FAIR data, code and workflows– maximises the scientific and societal impact and long-term value of our projects and is central to sustainability.
- The ICFA Data Lifecycle panel's best-practice recommendations for data preservation and open science:
  - were built on the expertise acquired by the HEP community.
  - show all stakeholders **their role** in **preserving our knowledge**.
- An assessment process is ongoing to raise awareness, share solutions and incentivise dialogue.
- The evolution of AI is changing how we work and affects, among other things, long-term data reusability.
  - How do we guide this change so that we maintain our skills, and "responsibility and scientific integrity stay ours!" (– Andreas Hoecker)?



Thank you!

Questions?
Discussion?

Presentation template from SlidesCarnival