# DPHEP and ICFA Data Lifecycle Panel

DPHEP Workshop - October 03, 2024

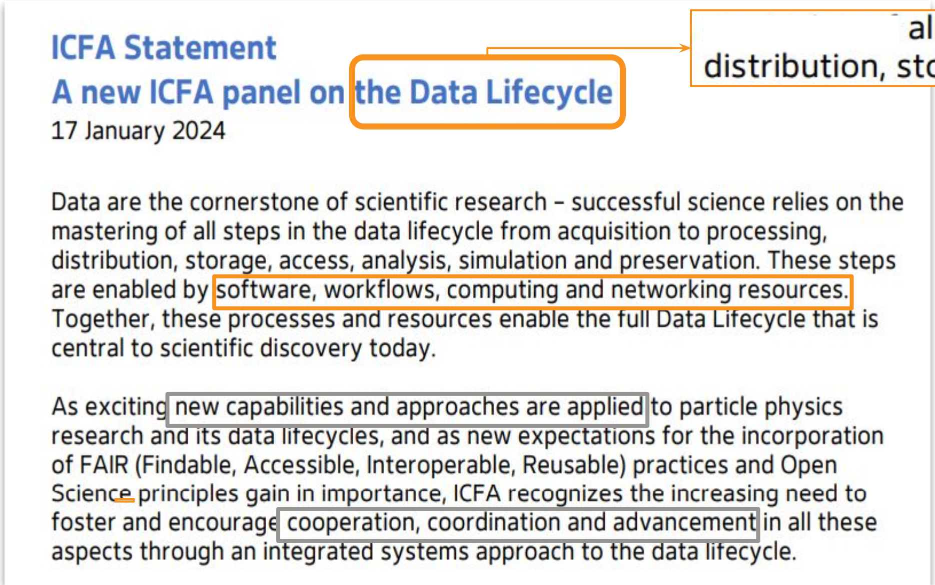
Kati Lassila-Perini
Helsinki Institute of Physics - Finland

# Data Lifecycle - What?

How actions on different systems at different points of Data Lifecycle affect the usability of data.

Preserved ripple marks from 2 Gy ago
Noitatunturi – Finland

## Is data preservation just the last step of the data lifecycle?



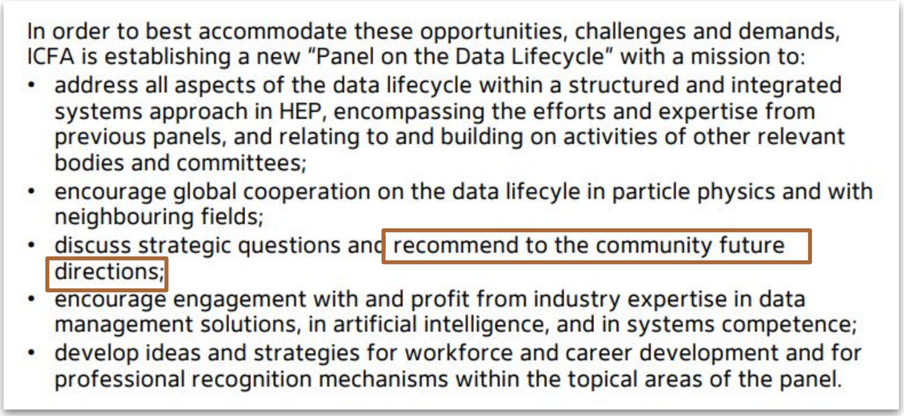
## ICFA Statement
## A new ICFA panel on the Data Lifecycle

17 January 2024

Data are the cornerstone of scientific research – successful science relies on the mastering of all steps in the data lifecycle from acquisition to processing, distribution, storage, access, analysis, simulation and preservation. These steps are enabled by software, workflows, computing and networking resources. Together, these processes and resources enable the full Data Lifecycle that is central to scientific discovery today.

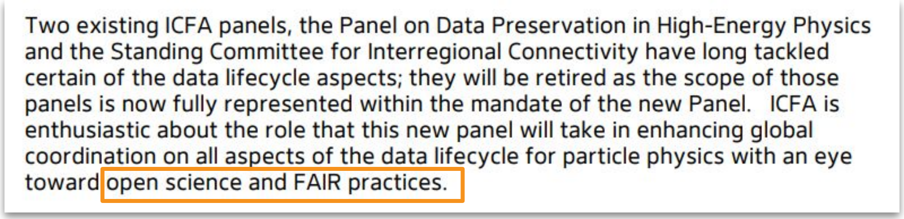
As exciting new capabilities and approaches are applied to particle physics research and its data lifecycles, and as new expectations for the incorporation of FAIR (Findable, Accessible, Interoperable, Reusable) practices and Open Science principles gain in importance, ICFA recognizes the increasing need to foster and encourage cooperation, coordination and advancement in all these aspects through an integrated systems approach to the data lifecycle.

all steps in the data lifecycle from acquisition to processing, distribution, storage, access, analysis, simulation and preservation.



In order to best accommodate these opportunities, challenges and demands, ICFA is establishing a new "Panel on the Data Lifecycle" with a mission to:

- address all aspects of the data lifecycle within a structured and integrated systems approach in HEP, encompassing the efforts and expertise from previous panels, and relating to and building on activities of other relevant bodies and committees;
- encourage global cooperation on the data lifecyle in particle physics and with neighbouring fields;
- discuss strategic questions and recommend to the community future directions;
- encourage engagement with and profit from industry expertise in data management solutions, in artificial intelligence, and in systems competence;
- develop ideas and strategies for workforce and career development and for professional recognition mechanisms within the topical areas of the panel.



Two existing ICFA panels, the Panel on Data Preservation in High-Energy Physics and the Standing Committee for Interregional Connectivity have long tackled certain of the data lifecycle aspects; they will be retired as the scope of those panels is now fully represented within the mandate of the new Panel. ICFA is enthusiastic about the role that this new panel will take in enhancing global coordination on all aspects of the data lifecycle for particle physics with an eye toward open science and FAIR practices.

**Actions on data affect their usability at different points of the lifecycle.**

We have heard great examples during the workshop!

## It is a cycle!

- Use of LEP/HERA data for future collider projects.
- Recovering CERNLIB enabling LEP data reuse.
- Phenix feeding their experience to EIC.
- Alice OD format benetting the current analysis.
- Atlas/CMS OD documentation benefitting collaboration members.
- Use of CMS OD in research.
- LHCb ntupling service.
- …

# Shout out!

- DP community is small, enthusiastic, active
  - Impressive progress again between the DPHEP workshops!
- Beyond these DPHEP workshops:
  - How could we reach out?
  - How could we reach “in” in the collaborations?
  - How do we attract young people?
- DP community has a solid expertise on
  - facilitating factors for data preservation and reuse
  - obstacles for data preservation and reuse
- Let’s share that expertise!

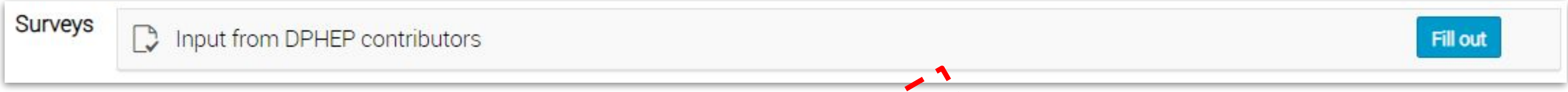
# Data Lifecycle panel

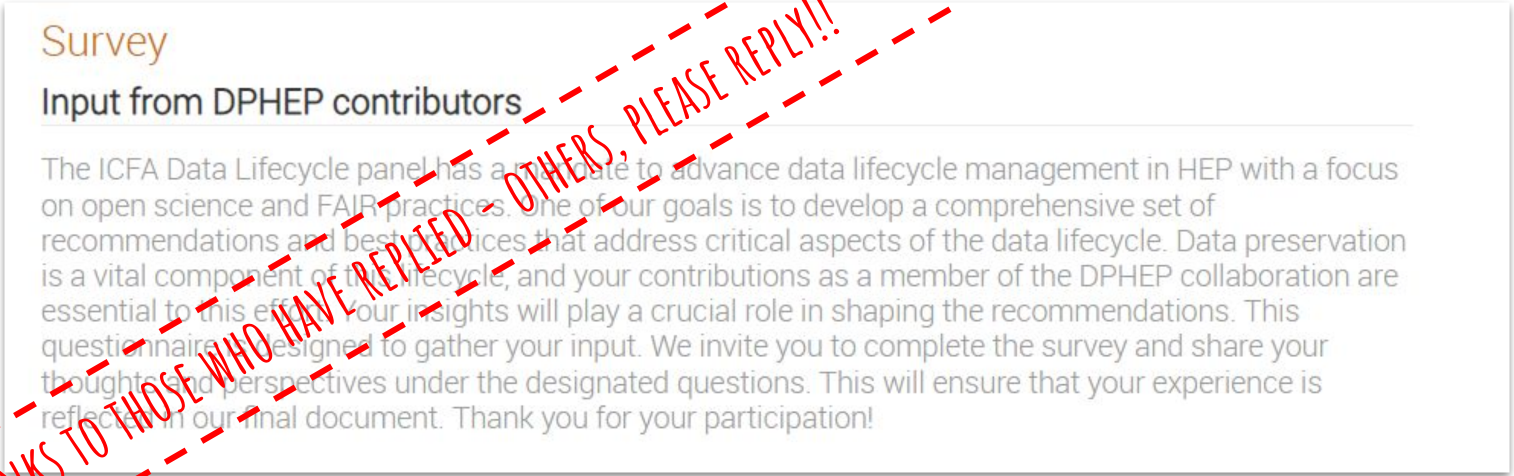
What can we do?

"... recommend to the community future directions ....

# Your input counts!

- Data lifecycle panel intends to develop recommendations for best practices to facilitate DP and data reuse
  - We want them to be concrete, specific and relevant to our domain.
  - We want them to be understandable to all stakeholders: from students and analysts to the experiment management.
- Therefore
  - Reaching out to *enablers* in our domain to hear their view:
    - DPHEP – here, right now!
    - HSF training organizers, trainers of AP skills in the experiments
  - Following the ongoing work for KPIs (Key Performance Indicators) for Open Science at CERN (and elsewhere?)
    - Recommendation and KPIs should match.





Survey

Input from DPHEP contributors

The ICFA Data Lifecycle panel has a mandate to advance data lifecycle management in HEP with a focus on open science and FAIR practices. One of our goals is to develop a comprehensive set of recommendations and best practices that address critical aspects of the data lifecycle. Data preservation is a vital component of this lifecycle, and your contributions as a member of the DPHEP collaboration are essential to this effort. Your insights will play a crucial role in shaping the recommendations. This questionnaire is designed to gather your input. We invite you to complete the survey and share your thoughts and perspectives under the designated questions. This will ensure that your experience is reflected in our final document. Thank you for your participation!

THANKS TO THOSE WHO HAVE REPLIED – OTHERS, PLEASE REPLY!!

*Your input counts!*

# So far: Facilitating factors

- Institutional support
- Technical solutions:
  - Reduced data formats
  - Open-source software
  - Software containers
  - Decoupling from specific environments
  - Data migration to more accessible storage
- Policy and collaboration:
  - LHC Open Data Policy
  - Best effort agreements with IT
  - Collaboration with CERN IT open data team
- Ongoing research needs:
  - Continued data analyses beyond main funding period
  - Regular publications from datasets
- Documentation and accessibility:
  - Full data provenance information
  - Dedicated tools for open data access
  - Clear instructions and easier processing
- Community factors:
  - Increased appreciation of preservation efforts
  - Positive feedback from the community
  - Small group of committed individuals
- Standardization:
  - Use of common packages and standard techniques
  - Central storage of experiment-specific software and documentation

# So far: Obstacles...

- Resource constraints:
  - Limited funding
  - Lack of dedicated person-power
  - Time constraints, especially at the end of analysis processes
- Policy and understanding issues:
  - Restrictive data access policies
  - Misunderstanding of data preservation vs. open data
  - Lack of awareness about preservation policies within experiments

- Technical challenges:
  - Proliferation of analysis frameworks
  - Complexity of analysis preservation
  - Software maintenance over long periods
  - Adapting to changes in computing infrastructure and OS support

- Documentation and standardization:
  - Sparse or fragmented documentation
  - Non-standardized recording of analysis information
  - Use of non-open formats for documentation
- Continuity and knowledge transfer:
  - Loss of human knowledge over time
  - Lack of continuity when individuals move on
  - Information stored in personal directories that may be deleted

# So far: ...Obstacles...

- Commitment and coordination:
  - Reliance on individual initiatives
  - Difficulty in uniting around a common vision
  - Weak language in policies leading to ambiguity

- Commitment and coordination:
  - Reliance on individual initiatives
  - Difficulty in uniting around a common vision
  - Weak language in policies leading to ambiguity

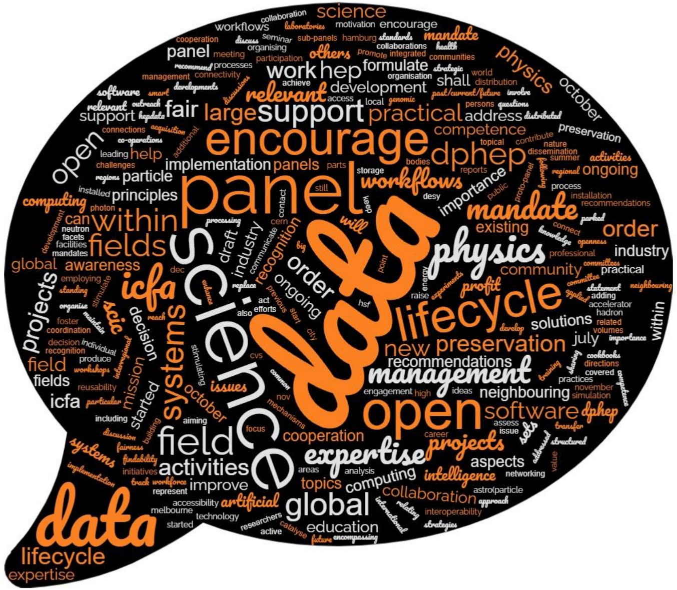
- Cultural factors:
  - Perception of low return on investment for preservation efforts
  - Pushback from various parties within experiments
  - Difficulty in openly discussing challenges across experiments

# Outlook

## Work through your input

Draft recommendations for best practices to facilitate data preservation and reuse.

Circulate for discussion within the DP/AP community.



# Thank you!

## *Questions?*

And thanks to SlidesCarnival for this free presentation template

*ICFA statement on the Data Lifecycle Panel*

*Mandate of the Data Lifecycle Panel*