

Input from open data contributors - DPHEP October 2024

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 section 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!

Survey

(This part can be implemented as a closed survey form)

Your role

Is your contribution towards data preservation(multiple choices)

- mandated by your experiment
- mandated by your institute
- your individual initiative

Is your work towards data preservation

- encouraged by your experiment
- supported by your experiment
- tolerated by your experiment
- encouraged by your institute
- supported by your institute
- tolerated by your institute
- or would you describe it differently:

What was your main motivation to get involved in the data preservation activities?

What are the main obstacles to getting involved in data preservation activities?

Activity

What fraction of your time have you dedicated to data preservation activities over the past two years?

- > 80%
- 50 - 80%
- 20 - 50%
- < 20%

Is this share

- fairly constant over the past two years
- fairly constant over a period longer than two years
- built up with shorter intense periods (in scale of weeks/months)

Do you plan to be involved in the data preservation activities in the future?

How many people are involved in data preservation activities in your experiment?

- 1
- 2-3
- 4-6
- 7-10
- > 10

Questions

Please provide your input with your name and experiment under the questions.

1 - Facilitating factors: What are/were the main factors facilitating data preservation in your experiment?

Kati Lassila-Perini CMS:

- early approval of the CMS open data policy
- dedicated group with a mandate defined in the policy
- resources from the host laboratory (open data storage, open data portal)
- somewhat well-defined reprocessing campaigns of legacy data and MC
- existence of the full data provenance information in the experiment-specific services
- open-source reconstruction and analysis software
- provision of software containers with experiment-specific SW
- work towards smaller data formats
- increased appreciation of the efforts

- ...

2 - Obstacles: What are the main obstacles to data preservation in the data management practices of your experiment? For “data management”, consider the management of data, software and workflows.

KLP CMS:

- open or preserved data are not yet flagged as such in many data management tools
- full data provenance is not readily available in the dataset metadata, it needs to be extracted from various experiment-specific services
- proliferation of analysis frameworks has made it challenging to provide usage instructions
- scale factors and corrections have been provided through various means (group-specific twiki pages, repositories) making it difficult to provide usage instructions
- analysis information, e.g. input datasets, code, is not recorded in a machine-readable way making it difficult to make it FAIR
- improvements in the data lifecycle chain are difficult to push through, in particular, if the impact is only at the data preservation phase
- many improvements are individual initiatives and continuity is often lost when people move
- lack of understanding of the effort needed for data releases
- ...

3 - Recommendations: Can you suggest recommendations for the management of data, software and workflows to facilitate data preservation?