Feedback from Data and Analysis Preservation, Recasting, and Reinterpretation white paper authors

Matthew Feickert
on behalf of the arXiv:2203.10057 white paper authors

University of Wisconsin-Madison

Snowmass CompF7 Session
July 21st, 2022

Snowmass white paper contribution

- Expressing ideas important to the authors of **Data and Analysis Preservation, Recasting, and Reinterpretation** Snowmass white paper
  - arXiv:2203.10057
- Focus on importance of long-term future reuse of published anlayses to maximise their scientific impact
  - Enable new analyses and reinterpretation of the results in the future
- Focus on collider physics (so too comments today)
  - Supportive of Cosmic Frontier recommenations

March 21, 2022

Submitted to the Proceedings of the US Community Study
on the Future of Particle Physics (Snowmass 2021)

Data and Analysis Preservation, Recasting, and Reinterpretation

TF07 (Collider Phenomenology in the Theory Frontier)
COMPF7 (Reinterpretation and long-term preservation of data and code)

Stephen Bailey[1], Christian Bierlich[2], Andy Buckley[3], Jon Butterworth[4], Kyle Cranmer[5], Matthew Feickert[6*], Lukas Heinrich[7], Axel Huebl[1], Sabine Kraml[8‡], Anders Kvellestad[9], Clemens Lange[10], Andre Lessa[11], Kati Lassila-Perini[12], Christine Nattrass[13], Mark S. Neubauer[6], Sezen Sekmen[14], Giordon Stark[15], Graeme Watt[16]

1 Lawrence Berkeley National Laboratory, USA **2** Lund University, Lund, Sweden **3** University of Glasgow, UK **4** University College London, UK **5** New York University, USA **6** University of Illinois at Urbana-Champaign, USA **7** Technische Universität München, Germany **8** Univ. Grenoble Alpes, CNRS, Grenoble INP, LPSC-IN2P3, Grenoble, France **9** University of Oslo, Norway **10** Paul Scherrer Institute, Villigen, Switzerland **11** Universidade Federal do ABC, Brazil **12** Helsinki Institute of Physics, Finland **13** University of Tennessee, Knoxville, USA **14** Kyungpook National University, Korea **15** SCIPP, UC Santa Cruz, CA, USA **16** IPPP, Durham University, UK

Corresponding authors:
* matthew.feickert@cern.ch, ‡ sabine.kraml@lpsc.in2p3.fr

Abstract

We make the case for the systematic, reliable preservation of event-wise data, derived data products, and executable analysis code. This preservation enables the analyses' long-term future reuse, in order to maximise the scientific impact of publicly funded particle-physics experiments. We cover the needs of both the experimental and theoretical particle physics communities, and outline the goals and benefits that are uniquely enabled by analysis recasting and reinterpretation. We also discuss technical challenges and infrastructure needs, as well as sociological challenges and changes, and give summary recommendations to the particle-physics community.

1

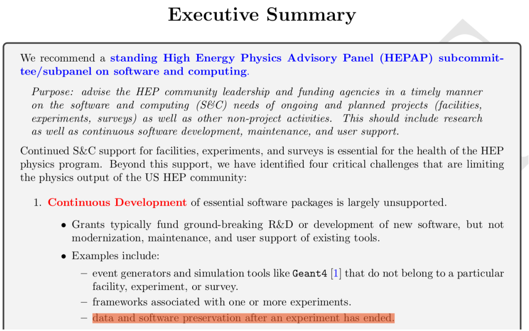
Highlights from CompF7 subgroup report and CompF report

- Thank you to the Computational Frontier and CompF7 conveners for writing these important reports!
- We strongly agree with and support the recommendations from the CompF7 subgroup report and for the CompF summary report **reinterpretation and long-term preservation** section as they are well aligned with our recommendations
  - Along with some suggestions to go further

CompF7 subgroup report recommendations

- **Recommendation 1**:
  Ensure that all current and future experiments have a **strategy and resourced program** for the long term preservation of data and analysis capabilities, including **beyond the lifetime of the individual experiments**.

- The legacy of modern experiments will be their analyses and the data will be used for decades to come. Critical to ensure this work is done.



Executive Summary

We recommend a **standing High Energy Physics Advisory Panel (HEPAP) subcommittee/subpanel on software and computing.**

Purpose: advise the HEP community leadership and funding agencies in a timely manner on the software and computing (S&C) needs of ongoing and planned projects (facilities, experiments, surveys) as well as other non-project activities. This should include research as well as continuous software development, maintenance, and user support.

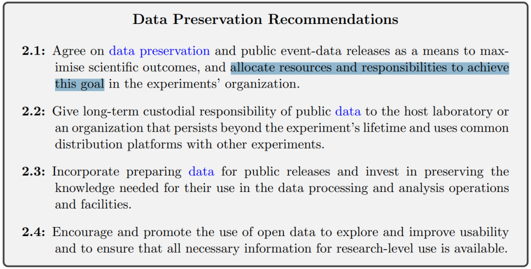
Continued S&C support for facilities, experiments, and surveys is essential for the health of the HEP physics program. Beyond this support, we have identified four critical challenges that are limiting the physics output of the US HEP community:

1. **Continuous Development** of essential software packages is largely unsupported.
   - Grants typically fund ground-breaking R&D or development of new software, but not modernization, maintenance, and user support of existing tools.
   - Examples include:
     - event generators and simulation tools like `Geant4` [1] that do not belong to a particular facility, experiment, or survey.
     - frameworks associated with one or more experiments.
     - data and software preservation after an experiment has ended.

Happy to see this as an executive summary item in the Computational Frontier report draft

CompF7 subgroup report recommendations

- **Recommendation 2**: Invest in **shared cyberinfrastructure** to preserve these data and support a comprehensive analysis from various experiments and surveys — both active and completed — in order to realize their full scientific impact. The infrastructure should support the **requisite theoretical inputs and computational requirements for analysis** as well as metadata and APIs to track provenance and incentivize participation.
- Examples: CERN Open Data Portal, HEPData, Zenodo



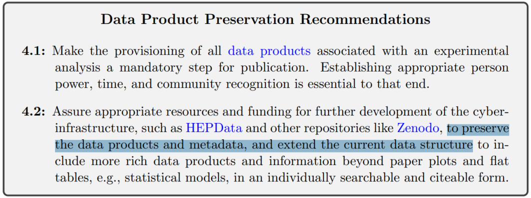
Data Preservation Recommendations

2.1: Agree on data preservation and public event-data releases as a means to maximise scientific outcomes, and allocate resources and responsibilities to achieve this goal in the experiments' organization.

2.2: Give long-term custodial responsibility of public data to the host laboratory or an organization that persists beyond the experiment's lifetime and uses common distribution platforms with other experiments.

2.3: Incorporate preparing data for public releases and invest in preserving the knowledge needed for their use in the data processing and analysis operations and facilities.

2.4: Encourage and promote the use of open data to explore and improve usability and to ensure that all necessary information for research-level use is available.



Data Product Preservation Recommendations

4.1: Make the provisioning of all data products associated with an experimental analysis a mandatory step for publication. Establishing appropriate person power, time, and community recognition is essential to that end.

4.2: Assure appropriate resources and funding for further development of the cyberinfrastructure, such as HEPData and other repositories like Zenodo, to preserve the data products and metadata, and extend the current data structure to include more rich data products and information beyond paper plots and flat tables, e.g., statistical models, in an individually searchable and citeable form.

arXiv:2203.10057

Going further: Explicit support

Strongly encourage explicitly noting that this work is not free in personel and funds in the CompF summary report (exists implicilty in the Executive Summary).

For these recommendations to be successful, there will need to be **dedicated support**:

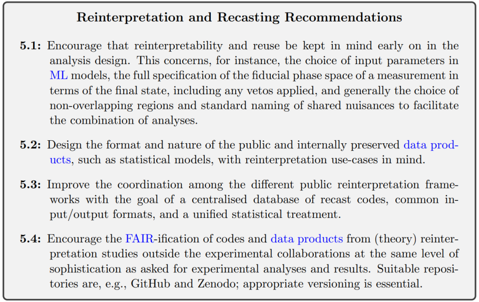
- People to work in these areas
- Provide the necessary infrastucture and services and associated maintainance

Happy to see this explicitly mentioned in the **CompF7 subgroup report Executive Summary**

- "US funding agencies should coordinate with international partners such as the CERN Open Data Portal and fund additional resources as needed to ensure that all US-supported projects have data and analysis preservation support, including post-operations and including non-collider programs."

Going further: Cross-field collaboration, reinterpretation, recasting

- Reinterpretation and recasting movivates multiple projects/goals acorss the field
  - Data and data product preservation: CERN Open Data Portal, HEPData, Zenodo
  - Tight information exchange loop between experiments and theory
  - FAIR-ification of software and data products
- "Designing and implementing datasets and analyses with this reuse in mind helps guide the pragmatic choices for where preservation effort is best spent." — CompF7 subgroup report
  - Stronger advocacy for reinterpretation and recasting helps motivate many of our other goals



Reinterpretation and Recasting Recommendations

5.1: Encourage that reinterpretability and reuse be kept in mind early on in the analysis design. This concerns, for instance, the choice of input parameters in ML models, the full specification of the fiducial phase space of a measurement in terms of the final state, including any vetos applied, and generally the choice of non-overlapping regions and standard naming of shared nuisances to facilitate the combination of analyses.

5.2: Design the format and nature of the public and internally preserved data products, such as statistical models, with reinterpretation use-cases in mind.

5.3: Improve the coordination among the different public reinterpretation frameworks with the goal of a centralised database of recast codes, common input/output formats, and a unified statistical treatment.

5.4: Encourage the FAIR-ification of codes and data products from (theory) reinterpretation studies outside the experimental collaborations at the same level of sophistication as asked for experimental analyses and results. Suitable repositories are, e.g., GitHub and Zenodo; appropriate versioning is essential.

arXiv:2203.10057

Challenges and Opportunities

Challenges

- Useful analysis preservation does take more work at the individual analysis level. To promote analysis preservation, the community needs to tangibly incentivize those doing the work.
  - “sociological” challenges in our white paper
- Funded infrastructure for preservation beyond the lifetime of experiments is limited.
  - “technical/infrastructure” challenges

Opportunities

- Broad support for concept of data and analysis preservation. Now is a time to push for this to recieve more support and buy-in.
- Improving the infrastructure around data and analysis presevation allows for richer and more complex data products to be preserved and used (e.g. DSL specs, full probability models).
- Well curated data and data products can be positive force for Equity, Diversity, and Inclusion initiatives

Summary

- **Strongly support** the current recommenations in the CompF7 subgroup report
- Happy to see that CompF7 goals have been noted in the Computational Frontier **summary report executive summary**
- Encourage the Computational Frontier summary report **to go further** and make **explicit requests for support** (funding, infrastructure, personel) and strengthen **support of reinterpretation and recasting** to CompF7 goals
- There will be challenges to move these recommendations forward, but as a community we can **leverage those into opportunities for the whole field**