Open is not enough

Xiaoli Chen, Sünje Dallmeier-Tiessen*, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, Diego Rodriguez Rodriguez, Tibor Šimko*, Tim Smith, Ana Trisovic, Anna Trzcinska, Ioannis Tsanaktsidis, Markus Zimmermann, Kyle Cranmer, Lukas Heinrich, Gordon Watts, Michael Hildreth, Lara Lloret Iglesias, Kati Lassila-Perini and Sebastian Neubert

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness.

Open science and reproducible research have become pervasive goals across research communities, political circles and funding bodies1–3. The understanding is that open and reproducible research practices enable scientific reuse, accelerating future projects and discoveries in any discipline. In the struggle to take concrete steps in pursuit of these aims there has been much discussion and awareness-raising, often accompanied by a push to make research products and scientific results open quickly.

Although these are laudable and necessary first steps, they are not sufficient to bring about the transformation that would allow us to reap the benefits of open and reproducible research. It is time to move beyond the rhetoric and the trust in quick fixes are not sufficient to bring about the transformation that would make research products and scientific results open quickly.

Discussion and awareness-raising, often accompanied by a push to take concrete steps in pursuit of these aims there has been much future projects and discoveries in any discipline. In the struggle to make research products and scientific results open quickly.

Thus, we argue that having the reuse of research results as a goal requires the adoption of new research practices during the data analysis process. Such practices need to be tailored to the needs of each given discipline with its particular research environment, culture and idiosyncrasies. Services and tools should be developed with the idea of meshing seamlessly with existing research procedures, encouraging the pursuit of reusability as a natural part of researchers’ daily work (Fig. 1). In this way, the generated research products are more likely to be useful when shared openly.

In tackling the challenge of enabling reusable research, we keep these ideas as our guiding light when putting changes into practice in our community—high-energy physics (HEP). Here, we illustrate our approach, particularly through our work at CERN, and present our community’s requirements and rationale. We hope that the explanation of our challenges and solutions will stimulate discussions around the practical implementation of workflows for reproducible and reusable research more widely in other scientific disciplines.

Approaching reproducibility and reuse in HEP

To set the stage for the rest of this piece, we first construct a more nuanced spectrum in which to place the various challenges facing HEP, allowing us to better frame our ambitions and solutions. We choose to build on the descriptions introduced by Carole Goble and Lorena A. Barba shown in Table 1.

These concepts assume a research environment in which multiple labs have the equipment necessary to duplicate an experiment, which essentially makes the experiments portable. In the particle physics context, however, the immense cost and complexity of the experimental set-up essentially make the independent and complete replication of HEP experiments unfeasible and unhelpful. HEP experiments are set up with unique capabilities, often being the only facility or instrument of their kind in the world; they are also constantly being upgraded to satisfy requirements for higher energy, precision and level of accuracy. The experiments at the Large Hadron Collider (LHC) are prominent examples. It is this uniqueness that makes the experimental data valuable for preservation so that it can be later reused with other measurements for comparison, confirmation or inspiration.

Our considerations here really begin after gathering the data. This means that we are more concerned with repeating or verifying the computational analysis performed over a given dataset rather than with data collection. Therefore, in Table 2 we present a nuanced spectrum of these definitions that takes into account a research environment in which ‘experimental set-up’ refers to the implementation of a computational analysis of a defined dataset, and a ‘lab’ can be thought of as an experimental collaboration or an analysis group.

In the case of computational processes, physics analyses themselves are intrinsically complex due to the large data volume and algorithms involved4. In addition, the analysts typically study more than one physics process and consider data collected under different running conditions. Although comprehensive documentation on the analysis methods is maintained, the complexity of the software implementations often hides minute but crucial details,
In absence of solutions for analysis capture and preservation, knowledge of specific methods and how they are applied to a given physics analysis might be lost. To tackle these community-specific challenges, a collaborative effort (coordinated by CERN, but involving the wider community) has emerged, initiating various projects, some of which are described below.

Reusable and openness. The HEP experimental collaborations operate independently of each other, and they do not share physics results until they have been rigorously verified by internal review processes. A physicist may perform further data reduction and selection procedures, which are followed by a statistical analysis on the data. Physics results are derived taking into account statistical and systematic uncertainties. The results often summarize which theoretical models have predictions that are consistent with the observations once background estimates have been included. The analysis assets being used by the individual researcher include the information about the collision and simulated datasets, the detector conditions, the analysis code, the computational environments, and the computational workflow steps used by the researcher to derive the histograms and the final plots as they appear in publications. Credit: CERN (a); Dave Gandy (b, c, code icon); SimpleIcon (b, c, gear icon); Andrian Valeanu (b, c, data icon); Umar Irshad (c, paper icon); Freepik (c, workflow icon).

Table 1 | Terminology related to reproducible research introduced by Carole Goble and Lorena A. Barba

| Term   | Purpose | Description |
|--------|---------|-------------|
| Rerun  | Robust  | Variations on experiment and set-up, conducted in the same lab |
| Repeat | Defend  | Same experiment, same set-up, same lab |
| Replicate | Certify   | Same experiment, same set-up, independent lab |
| Reproduce | Compare | Variations on experiment and set-up, independent labs |
| Reuse  | Transfer| Different experiment |
However, it is necessary to ensure the usability of the research in the long term. This is particularly challenging today, as much of the analysis code is available primarily within the small team that performs an analysis. We think that reproducibility requires a level of attention and care that is not satisfied by simply posting undocumented code or making data ‘available on request’.

In the particular case of particle physics, it may even be true that openness itself, in the sense of unfettered access to data by the general public, is not necessarily a prerequisite for the reproducibility of the research. Take the LHC collaborations as an example: while they generally strive to be open and transparent in both their research and their software development, analysis procedures and the previously described challenges of scale and data complexity mean that there are certain necessary reproducibility use cases that are better served by a tailored tool rather than an open data repository.

Such tools need to preserve the expertise of a large collaboration that flows into each analysis. Providing a central place where the disparate components of an analysis can be aggregated at the start, and then evolve as the analysis gets validated and verified, will fill this valuable role in the community. Confidentiality might aid this process so that the experts can share and discuss in a protected space before successively opening up the content of scrutiny to ever larger audiences, first within the collaboration and then later via peer review to the whole HEP community.

Cases in point are the CERN Analysis Preservation (CAP) and Reusable Analyses (REANA), which will be described in more detail below. Their key feature is that they leave the decision as to when a dataset or a complete analysis is shared publicly in the hands of the researchers. Open access can be supported, but the architecture does not depend on either data or code being publicly available.

This gives the experimental collaborations full control over the release procedure and thus fully supports internal processing, review protocols and possible embargo periods. Hence, the service is accessible to the thousands of researchers who need the information it contains in order to replicate or reuse results, but the public-facing functions in HEP are better served by other services, such as CERN Open Data, HEPData and INSPIRE.

The standard data deluge in particle physics is another challenge that calls for separate approaches for reproducibility, reusability and openness. As we do not have the computational resources to enable open access and processing of raw data, there needs to be a decision on the level at which the data can meaningfully be made open to allow valuable scrutiny by the public. This is governed by the individual experiments and their respective data policies.

### Enabling open and reusable research at CERN

The CERN Analysis Preservation and reuse framework consists of a set of services and tools, sketched in Fig. 1, that assist researchers in describing and preserving all the components of a physics analysis such as data, software and computing environment—addressing the points discussed earlier. These, along with the associated documentation, are kept in one place so that the analysis, or parts of it, can be reused even several years after the publication of the original scientific results.

The CERN Analysis Preservation and reuse framework relies on three pillars:

1. **Describe**: adequately describe and structure the knowledge behind a physics analysis in view of its future reuse. Describe all the assets of an analysis and track data provenance. Ensure sufficient documentation and capture associated links.
2. Capture: store information about the analysis input data, the analysis code and its dependencies, the runtime computational environment and the analysis workflow steps, and any other necessary dependencies in a trusted digital repository.

3. Reuse: instantiate preserved analysis assets and computational workflows on the compute clouds to allow their validation or execution with new sets of parameters to test new hypotheses.

All of these services, developed through free and open source software, strive to enable FAIR compliant data\textsuperscript{20} and can be set up for other communities as they are implemented using flexible data models. For all these services, capturing and preserving data provenance has been a key design feature. Data provenance facilitates reproducibility and data sharing as it provides a formal model for describing published results\textsuperscript{7}. CERN Analysis Preservation. The CERN Analysis Preservation (CAP) service is a digital repository instance dedicated to describing and capturing analysis assets. The service uses a flexible metadata structure conforming to JavaScript Open Notation (JSON)
schematics that describe the analysis in order to help researchers identify, preserve and find the information about components of analyses. These JSON components define everything from experimental configurations to data samples and from analysis code to links to presentations and publications. By assembling such schematics, we are creating a standard way to describe and document an analysis in order to facilitate its discoverability and reproducibility.

The CAP service features a ‘push’ protocol that enables individual researchers to deposit material either by means of a user interface or with an automated command-line client. In the case of primary data, it can store links to data deposited in trusted long-term preservation stores used by the HEP experiments. For software and intermediate datasets, it can also completely ingest the material referenced by the researcher.

The CAP service can also ‘pull’ information from internal databases of LHC collaborations, when such information exists. Aggregating various sources of information from existing databases, source code repositories and data stores is an essential feature of the CAP service, helping researchers find and manage all the necessary information in a central place. Such an aggregation and standardization of data analysis information offers advanced search capabilities to researchers, facilitating discovery and search of high-level physics information associated with individual physics analyses.

**REANA.** We argue that physics analyses ideally should be automated from inception in such a way that they can be executed with a single command. Automating the whole analysis while it is still in its active phase permits to both easily run the ‘live’ analysis process on demand as well as to preserve it completely and seamlessly once it is over and the results are ready for publication. Thinking of restructuring a finished analysis for eventual reuse after its publication is often too late. Facilitating future reuse starts with the first commit of the analysis code.

This is the purpose served by the Reusable Analyses service, REANA: a standalone component of the framework dedicated to instantiating preserved research data analyses on the cloud. While REANA was born from the need to rerun analyses preserved in the CERN Analysis Preservation framework, it can be used to run ‘active’ analyses before they are published and preserved.

Using information about the input datasets, the computational environment, the software framework, the analysis code and the analysis configuration, REANA leverages modern container technologies to encapsulate the runtime environment necessary for various analysis steps. REANA supports several different container technologies (Docker\(^1\), Singularity\(^2\)), compute clouds (Kubernetes\(^3\)/OpenShift\(^4\), HTCondor\(^5\)), shared storage systems (Ceph\(^6\), EOS\(^7\)) and structured workflow specifications (CWL\(^8\), Yadage\(^9\)) as they are used in various research groups.

**RECAST.** RECAST\(^{10}\) is a notable example of an application built around reusable workflows, which targets a specific particle physics use case. In particular, RECAST provides a gateway to test alternative computational workflows and new physical theories by simulating what those theories predict and then running the simulated data through the analysis workflow used for a previous publication. The application programming interface exposes a restricted class of trustworthy, high-impact queries on the data. The experiment's data and the data processing workflow need not be exposed directly. Furthermore, the experimental collaborations can optionally maintain an approval process for the new result.
The system has been used internally to streamline the reinterpretation of several experiments, and ultimately could be opened to independent researchers outside of the LHC collaborations.

CERN Open Data. The CERN Open Data portal was released in 2014 amid a discussion as to whether the primary particle physics data, due to its large volume and complexity, would find any use outside of the LHC collaborations. In 2017, Thaler and colleagues confirmed their jet substructure model predictions using the official CMS Higgs-to-four-lepton analysis results (Fig. 3). This shows that the CERN Open Data service fulfils a different and complementary use case to the CERN Analysis Preservation framework. The openness alone does not sufficiently address all the required use cases for reusable research in particle physics that is naturally born “closed” in experimental collaborations before the analyses and data become openly published.

Challenging, but possible

In this paper we have discussed how open sharing enables certain types of data and software reuse, arguing that simple compliance...
with openness is not sufficient to foster reuse and reproducibility in particle physics. Sharing data is not enough; it is also essential to capture the structured information about the research data analysis workflows and processes to ensure the usability and longevity of results.

Research communities may start by using open data policies and initiating dialogues on data sharing, while embracing the reproducibility and reuse principles early on in the daily research processes. We compiled a few guiding principles that could support such dialogues (Box 1). In particle physics, the possibility of actual internal or external reuse of research outputs is an intrinsic motivation for taking part in these activities; one could assume the same for many other scientific communities.

Using computing technologies available today, solving the challenges of open sharing, reproducibility and reuse seems more feasible than ever, helping to keep research results viable and reusable in the future.

References

Received: 6 April 2018; Accepted: 10 September 2018; Published online: 15 November 2018

Acknowledgements

Considerations and services described here benefit from funding and discussions from the EC funded projects FREYA (777523), THOR (654039) and CRISP (283745). Furthermore, we acknowledge support for the DASPOS (https://daspos.crc.nd.edu/) and Diana/HEP projects (http://diana-hep.org/).

Additional information

Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to S.D. or T.S. or A.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2019

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.