CERN Analysis Preservation: A Novel Digital Library Service to Enable Reusable and Reproducible Research

Xiaoli Chen[1,5], Sünje Dallmeier-Tiessen[1(✉)], Anxhela Dani[1,2], Robin Dasler[1], Javier Delgado Fernández[1,4], Pamfilos Fokianos[1], Patricia Herterich[1,3], and Tibor Šimko[1]

[1] CERN, Geneva, Switzerland
{xiaoli.chen,sunje.dallmeier-tiessen,anxhela.dani,robin.dasler,
javier.delgado.fernandez,pamfilos.fokianos,
patricia.herterich,tibor.simko}@cern.ch
[2] Alexander Technological Educational Institute of Thessaloniki, Thessaloniki, Greece
[3] Humboldt-Universität zu Berlin, Berlin, Germany
[4] Universidad de Oviedo, Oviedo, Spain
[5] University of Sheffield, Sheffield, UK

Abstract. The latest policy developments require immediate action for data preservation, as well as reproducible and Open Science. To address this, an unprecedented digital library service is presented to enable the High-Energy Physics community to preserve and share their research objects (such as data, code, documentation, notes) throughout their research process. While facing the challenges of a "big data" community, the internal service builds on existing internal databases to make the process as easy and intrinsic as possible for researchers. Given the "work in progress" nature of the objects preserved, versioning is supported. It is expected that the service will not only facilitate better preservation techniques in the community, but will foremost make collaborative research easier as detailed metadata and novel retrieval functionality provide better access to ongoing works. This new type of e-infrastructure, fully integrated into the research workflow, could help in fostering Open Science practices across disciplines.

Keywords: Research data · Long-term preservation · Reproducible science · Digital repository · Research workflow

1 Introduction

In response to the pressing demand for scientific output to be reproducible and re-usable, as well as the call for science to generate wider societal impact, an increasing number of funding agencies and research organizations have started

The original version of this chapter was revised: The Fig. 2 was corrected. The erratum to this chapter is available at 10.1007/978-3-319-43997-6_45

N. Fuhr et al. (Eds.): TPDL 2016, LNCS 9819, pp. 347–356, 2016.
DOI: 10.1007/978-3-319-43997-6_27

to mandate better data preservation practices and to call for the opening of data at varying levels and time intervals, with respect to disciplinary conventions [1].

Preservation and sharing of research results are the keys to scientific progress, and this fact has long been recognized within the high-energy physics (HEP) discipline as evidenced by its collaborative nature and well established preprint culture [2]. In order to accelerate communication, manuscripts have been shared among the HEP community prior to publication for many years, first by mail and more recently via open digital repositories after the invention of the World Wide Web at CERN.

Coordinated data preservation practices and research data sharing for HEP, on the other hand, have just gained momentum in recent years [3]. HEP experiments, which can involve from hundreds to thousands of collaborators, produce large datasets that undergo intricate quality assurance processes before entering analysis phase, leaving a trail of research outputs of different refinement levels and usage [4]. Data are available within the collaborations for analyses, and only the final results, usually appearing in the form of plots and tables, are included in the final publication. In 2014, after complex preparation, CERN launched its Open Data Portal to address the matter of data sharing [5]. The platform made preserved datasets and accompanying tools accessible to both the general public and the research community, yet through it emerged new challenges in research provenance preservation, which is essential for analyses reusability, reproducibilty and discoverability. Hence, the inception of an additional service, CERN Analysis Preservation.

CERN Analysis Preservation (CAP) is a digital library service designed with HEP's unique disciplinary research workflow in mind, aimed at capturing the research data analysis workflow steps and resulting digital objects. Emphasis is given to the contextual knowledge needed to reproduce an analysis. Hence, CAP can be seen as a concrete step towards better reuse of unique research materials and as a way to facilitate future reproducibility of results.

2 Concept

In contrary to traditional approaches, where documentation and preservation happen in separate formats after the completion of an analysis, CAP provides a centralized platform on which scientists can document their analysis as early as the start of a new project. In addition, scientists can keep track of any aspect or step of an analysis as well as related research objects within their collaboration.

Researchers can submit their content, datasets, code, intermediate documentation of processing steps, quality assurance processes or internal notes. The tool is connected to the collaborations' databases, which enables auto-completion of many parts of the analysis metadata. The service assists researchers in preserving implicit knowledge and facilitates sharing of and access to research materials for future use. Alongside the analysis details and processing steps, reusable information for the analysis and supplementary documents are also captured.

From the outset of its design and development, the CAP team at CERN has been working closely with the physics collaborations to deliver a tailour-made

solution for the HEP community. Together with the community, the service providers in the Information Technology Department and the Scientific Information Service at CERN framed the use cases [6]. CAP aims to document and preserve tasks that traditionally are implicit to the research process - like personal log-book keeping for an ongoing analysis, exchange of preliminary results, quality assurance processes, links from an analysis to approval or publication processes - all in an integrated fashion. By supporting the physics analysis work flow electronically and centrally, CAP mitigates some long standing pains in the daily work, such as parallel versions, additional approval procedures, conflicted results resolution, and analyses combination or reproduction.

Overall this new digital library service helps researchers manage their research workflows better, by helping internal works and data become more accessible and findable despite being in a large collaboration. This will save time and effort on the researcher side, present collaborators with the chance to build on other peoples work more quickly, as well as offer new personnel a better chance to familiarise themselves with ongoing works in large scale collaborations.

From the service providers side, this new e-infrastructure service helps establishing enhanced data curation and standardization workflows. This approach is considered a milestone to facilitate reusable science, i.e. to use data or rerun an analysis in many years to come. With the unprecedented granularity, CAP provides a new lens for HEP scientists to look at on-going and past analyses. Further more, through its association with the CERN Open Data Portal, INSPIRE and CDS, CAP significantly lowers the barriers for HEP scientists to engage in data reuse and eventual reproducible analysis, and hence fosters Open Science practices.

3 Technology

The CAP service is built on top of the Invenio digital library framework [7]. Invenio provides an ecosystem of standalone independent packages that permit to build a custom digital repository solution oriented towards various use cases, such as an integrated library system, a digital document repository, a multimedia archive, or a data repository. In the context of CAP, the Invenio framework has been used and extended with several large-data oriented features.

The managed data are modelled in JSON format that is conforming to certain JSON Schema to ensure the compliance of captured JSON snippets with standard metadata requirements. The platform draws inspiration from the Open Archival Information System (OAIS) recommended practices to ensure long-term preservation of captured assets. The JSON snippets are stored in the Invenio digital repository database and sent to an Elasticsearch cluster that is being used for indexing and information retrieval needs.

CAP aims at preserving the complete environment. This necessitates to build connectors to various tools used in physics analyses in order to be able to take a consistent snapshot of the analysis process for knowledge capture and preservation.

The overall architecture (Fig. 1) of the CAP platform connects to various tools:

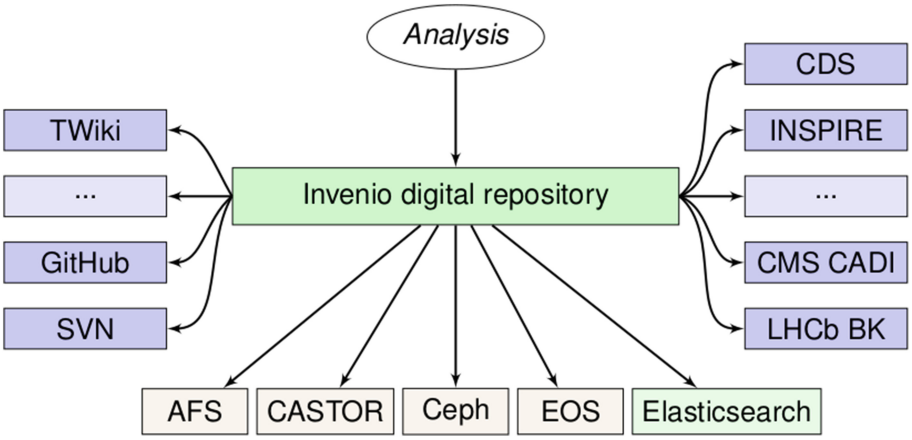

Fig. 1. The architecture overview of the CERN Analysis Preservation platform

The system notably includes:

- the "core" of the Invenio digital repository platform itself;
- the flexible file storage abstraction layer using the EOS storage backend;
- the connectors to harvest final datasets from various storage systems used by individual physicists, such as CASTOR;
- the connectors to harvest user code from Git and Subversion repositories used to develop the analysis code;
- the connectors to documentation systems such as TWiki are also being developed;
- the platform may automatically ingest information from internal collaboration systems (e.g. CMS CADI);
- for the approved open data, the platform may push parts of information to publishing platforms such as (CDS and INSPIRE).

In order to facilitate the reproduction of an analysis even many years after its initial publication, CAP aims at instantiating the original analysis environment on the CERN OpenStack cloud by means of technologies such as CernVM virtualisation or container-based solutions using Docker, Jupyter or Everware.

4 Functionality and UI

The service offers various entry points and functionalities to suit various users, so that the tool will more easily become an intrinsic part of the research process. Researchers can use a submission form (Figs. 2 and 3 show parts of an example) to submit their content, or they can select a more automated means to do so (via the REST API). Based on the flexible JSON schemata tailoured input formats can be offered for experiments and working groups. Auto-complete functionality

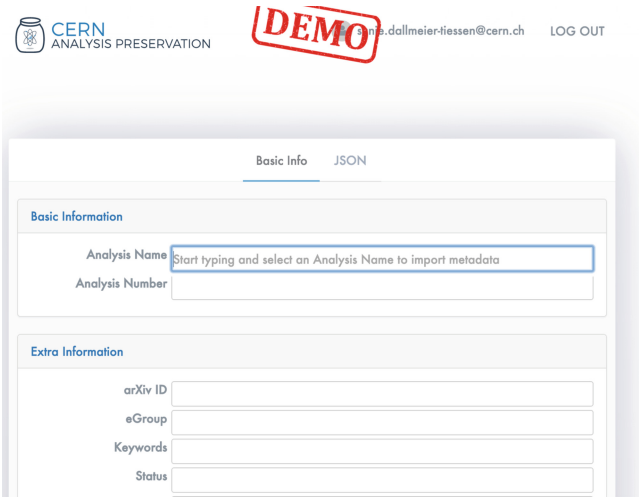

Fig. 2. First part of a submission form for one experiment. Auto-complete functionality (using information from a range of existing databases) allows researchers to easily fill in the form without much extra effort.

for the submission form (based on connections made to existing databases within the collaboration) should reduce the time researchers need to spend on using this e-infrastructure service for submitting their content.

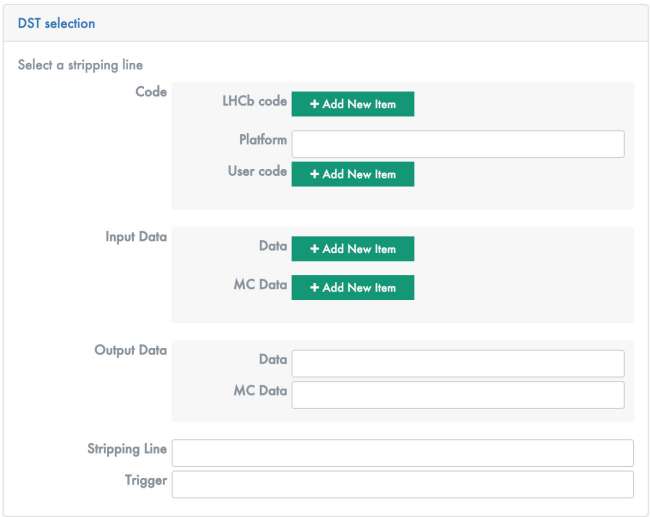

Fig. 3. Second part of a submission form for one experiment. It contains links to code, data and physics information.

By allowing for the inclusion of such diverse content and knowledge, the service offers a central place for information about an analysis. This knowledge could help to ease the analysis and publication approval procedures that take place within the collaborations. When researchers submit an analysis to the dedicated board for review, they could compile the required information from the tool very easily, so that there is no need to gather the information from disparate, unconnected sources as per the traditional work flow. This functionality is under discussion, but not yet implemented.

Most important for the advancement of science is the capture of all the essential information used in an analysis in a central place (CAP), which permits advanced retrieval opportunities and knowledge sharing. Before, information about an analysis was scattered around many databases. With CAP users will be able to find an analysis with specific parameters, processed with a specific algorithm, or using a specific dataset or simulation (just to name a few examples). This opens ups new possibilities for internal collaboration.

CAP enables researchers to assign permissions to their submissions, so as to allow individuals control over the privacy of their submission while accounting for the collaborative nature of the discipline. The submitter can invite a group to view and/or edit the submission. A permissions tab allows the specification of each invitee's rights (Fig. 4). Given the work in progress nature of the content in CAP, this functionality has been considered crucial by the community.

While Analysis Preservation is considered an internal digital library service, it will offer open publishing options via partner portals. The researcher himself needs to trigger this process and content should be approved internally prior to publication. A DOI registration workflow for such content makes sure that reuse of materials can be tracked and credit can be attributed to the contributors.

Based on requests from the community, the service will be prepared to facilitate computational workflows by allowing for the re-use of analysis data and

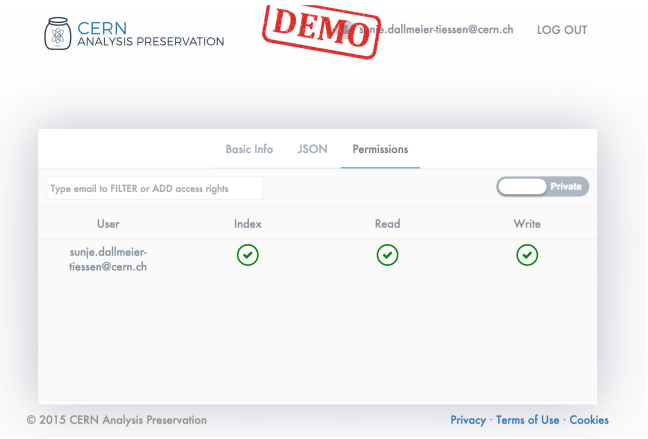

Fig. 4. Permissions setting for an analysis record on CAP. This enables researchers to share their analysis internally depending on their needs.

materials in an automated fashion. Additionally, it is under investigation how Jupyter notebooks might be integrated to accommodate and facilitate modern research practices in support of open science goals.

5 Content and Metadata

The service providers, together with preservation managers from the collaborations, identified the core components of a physics analysis and the metadata therein. This is very specific to the individual collaboration and, in many cases, even within the individual working groups of a collaboration. The tailoured JSON schemata have been developed with the community to make sure researchers can submit all content that is relevant for their analysis (the JSON schema then translates to the submission form; an example excerpt for the LHCb experiment can be seen in Fig. 5). This approach should make sure sufficient information is in place for better internal accessibility, future reuse and reproducibility.

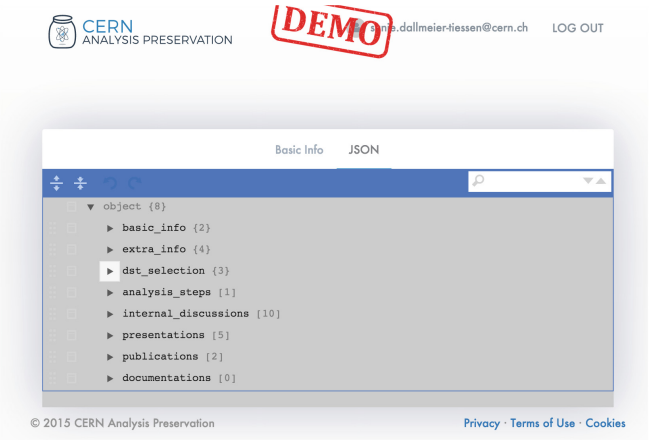

Fig. 5. An example of the high-level JSON components for a physics analysis of the LHCb experiment.

The service is connected to the core databases of the partnering collaborations. At the moment the service has access via APIs to CMS and LHCb databases. Work is in progress to integrate with other collaboration's tools as well. Harvesting databases facilitates auto-completion for most fields of the analysis submission form. Connecting to the existing databases within CMS, for example, provides basic information, such as authors and title, based on the unique ID of an analysis. Further important information about the ongoing analysis includes the operating system, the analysis software, the code, input data files and the output data. By following the ID in the CMS internal databases, one can retrieve information such as internal discussions, presentations and publications related

to an analysis when they are available. Also, detailed physics information, which is needed for future reuse, can be available: e.g. details about the primary and simulated (Monte Carlo) datasets; physics information that refers to the final state particles, cuts and vetos.

Likewise in LHCb's databases the service harvests all the needed content and displays basic information like the distinctive analysis name and number. Additional information is provided about the responsible working groups and the status of an analysis. Furthermore, the tool gathers valuable components of an analysis, such as the code, the input and output data as well as their simulated (Monte Carlo) data. In the future, collaborations' Twiki pages should also be integrated (as preserved snapshots), as they are actively used within the collaborations for internal discussions. Lastly, whenever available, the tool incorporates information related to presentations, publications and any other documentations of an analysis. All these connections should help researchers in the submission process. They can open the tool and enter the unique internal ID for analysis so that much of the information is auto-completed. Then, they only need to add the remaining pieces of information.

This iterative analysis processes throughout the research process and the collaborative preservation and editing of content among users will result in different versions of an analysis, i.e. improved or upgraded content. Hence, versioning is considered a very necessary component for CAP users and will be facilitated. The same applies to the documentation, which starts at the beginning of an analysis and that needs to reflect the changes, additions, and deletions throughout the entire process. To keep track of the changes the service would maintain explicit versions of the content.

High metadata quality is an integral part of the CAP system. The information that accompanies the code and data of an analysis for purposes of description, administration, technical functionality, use and preservation, assures the long-term usage. CAP enables comprehensive incorporation of metadata components specific to an individual analysis. This approach ensures a rich metadata source that can be utilized as open data publication demands emerge. In such cases, it would be mapped to standards like the Datacite Metadata Schema [8] to enable the publishing of complete and citable analysis records.

High quality metadata makes the data more visible, accessible and usable in the long term and is considered a key benefit for researchers who make use of the tool's search function. Complete and interoperable metadata are considered crucial components for future reproducibility. At this point it should be noted that CAP is not an effort to enforce a standard across experiments, which is being done in other disciplines, but it instead works flexibly with each collaboration's requirements. The overall similarity of analysis components allows for mapping between them, rather than globally enforced requirements. To address the standardization challenge in the long term, work is underway, together with the DASPOS project, to establish an ontology for the main analysis building blocks.

6 Conclusions and Next Steps

CERN Analysis Preservation puts the concept of digital library services into new use. By being embedded into the actual research workflow, research outputs throughout the analysis process can be preserved and shared internally. This creates new challenge in terms of (frequent) versioning and handling of very high data volumes. An analysis can be up to 10TB in volume currently and is expected to grow further over the course of the next few years. Overall, the main goal of CERN Analysis Preservation is to facilitate the reusability and reproducibility of an analysis even many years after its initial start or publication. To do so, partnerships are created with other tool providers that enable researchers to easily reuse, validate and execute preserved content.

For the future it is needed to extend the first prototype that is available, and integrate it even more with existing databases and processes of the experiments. This will be a key factor to foster adoption. Furthermore, the second phase of the project will begin considering the possibility of connections to analysis environments, in order to enable researchers to embed the discovered content easily into Jupyter notebooks and other tools.

The Open Source developing approach enables other disciplines to reuse the concept and tool. In particular, service providers for data-intensive disciplines might be interested to build service components flexible enough to serve a fast paced research environment with diverse, dynamic and possibly large scale research objects and better documentation.

Acknowledgements. CERN Analysis Preservation builds on the collaborative work from various teams and experiments. We are thankful for all the hard work that the LHC experiments and the Invenio team have been putting into this service so far.

References

1. European Commission. Open data: an engine for innovation, growth and transparent governance, December 2011. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF
2. Gentil-Beccot, A., Mele, S., Brooks, T.C.: Citing and reading behaviours in high-energy physics. Scientometrics **84**(2), 345–355 (2009)

3. Kogler, R., South, D.M., Steder, M.: Data preservation in high energy physics. J. Phys.: Conf. Ser. **368**, 012026 (2012)
4. Herterich, P., Dallmeier-Tiessen, S.: Data citation services in the high-energy physics community. D-Lib Mag. **22**(1/2) (2016)
5. Cowton, J., Dallmeier-Tiessen, S., Fokianos, P., Rueda, L., Herterich, P., Kunčar, J., Šimko, T., Smith, T.: Open data and data analysis preservation services for LHC experiments. J. Phys.: Conf. Ser. **664**(3), 032030 (2015)
6. Dallmeier Tiessen, S., Herterich, P., Igo-Kemenes, P., Šimko, T., Smith, T.: CERN analysis preservation (CAP) - Use Cases (2015). http://dx.doi.org/10.5281/zenodo.33693
7. Kunčar, J., Nielsen, L., Šimko, T.: Invenio v2.0: a pythonic framework for large-scale digital libraries, June 2014. http://urn.fi/URN:NBN:fi-fe2014070432294
8. DataCite Metadata Schema for the publication and Citation of Research Data, DataCite Metadata Working Group (2014). https://doi.org/10.5438/0011