EPJ Web of Conferences **245**, 06011 (2020) https://doi.org/10.1051/epjconf/202024506011
CHEP 2019

CERN Analysis Preservation and Reuse Framework: FAIR research data services for LHC experiments

Pamfilos Fokianos[1,*], *Sebastian* Feger[1], *Ilias* Koutsakis[1], *Artemis* Lavasa[1], *Rokas* Maciulaitis[1], *Kamran* Naim[1], *Jan* Okraska[1], *Antonios* Papadopoulos[1], *Diego* Rodríguez[1], *Tibor* Šimko[1], *Anna* Trzcinska[1], *Ioannis* Tsanaktsidis[1], and *Stephanie* van de Sandt[1]

[1]European Laboratory for Particle Physics, CERN, Geneva, Switzerland

Abstract. In this paper we present the CERN Analysis Preservation service as a FAIR (Findable, Accessible, Interoperable and Reusable) research data preservation repository platform for LHC experiments. The CERN Analysis Preservation repository allows LHC collaborations to deposit and share the structured information about analyses as well as to capture the individual data assets associated to the analysis. We describe the typical data ingestion pipelines, through which an individual physicist can preserve and share their final n-tuples, ROOT macros, Jupyter notebooks, or even their full analysis workflow code and any intermediate datasets of interest for preservation within the restricted context of experimental collaboration. We discuss the importance of annotating the deposited content with high-level structured information about physics concepts in order to promote information discovery and knowledge sharing inside the collaboration. Finally, we describe techniques used to facilitate the reusability of preserved data assets by capturing and re-executing reproducible recipes and computational workflows using the REANA Reusable Analysis platform.

1 Introduction

The preservation, management and accessibility of research data is an important challenge, one which a wide range of scientific disciplines have attempted to tackle. A variety of agencies and organisations have declared the importance of **Open Science** initiatives [1], and the European Union in particular has repeatedly argued for the value of **Open Data** [2] and their accessibility. In **High Energy Physics (HEP)** the culture of sharing notes, data, and discoveries (even before their publication) has been a staple for many years now, something evident in the creation of platforms like ArXiv, and even the creation of the World Wide Web.

CERN hosts the **Large Hadron Collider (LHC)**, the world's largest scientific instrument, producing millions of petabytes of data every year. All this information is securely stored, however, it is still very difficult to retrieve it at will, or to optimally reproduce an experiment using data and methodologies from many years ago. To address this challenge, a number of platforms and tools are either implemented at CERN, or created by teams that are heavily engaged with CERN.

CERN Analysis Preservation (CAP) started in 2014 in order to provide an integrated platform for scientific analysis, preservation and reuse. It is a digital repository for the description and preservation of all individual analysis assets, operating in the context of Open

*e-mail: pamfilos.fokianos@cern.ch

Science [3]. Through CAP, researchers can capture all the necessary metadata, save analysis components, engage in best practices for scientific preservation, and allow extensive collaboration between the researchers. This paper presents CERN Analysis Preservation as an end-to-end solution for scientific preservation and collaboration, with HEP-specific functionality. CAP aims at responding to two parallel demands:

Inside of the Collaborations: Due to the high degree of complexity of the analyses, preserving them becomes very challenging.

Externally: As a matter of policy, research funders are increasingly demanding that funded research programs institute data management frameworks to support data and knowledge preservation, but also the reuse, reinterpretation and reproducibility of research findings [4].

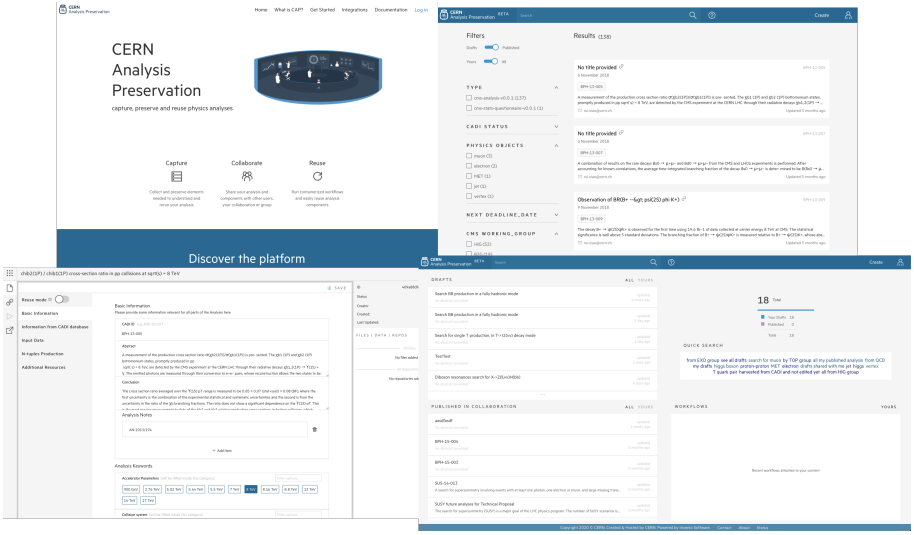

Figure 1: The CERN Analysis Preservation web interface.

2 Concept & Implementation

2.1 Motivation

CERN Analysis Preservation aspires to become the bridge between the different facets of a complete scientific analysis, i.e. a platform that can preserve all the different components, necessary to understand and rerun a HEP experimental analysis even several years in the future.

HEP researchers require a very specific workflow structure, with strictly defined policies on the access and preservation of their data, which all of the LHC collaborations follow. According to the **Study Group for Data Preservation and Long Term Analysis in High Energy Physics**, the previously mentioned policies state [5]:

1. all data (primary or otherwise) are to be available Open Access,

2. constructed datasets (simplified) are made available in educational and outreach contexts,

3. simulated datasets, including the necessary software are available Open Access after a predetermined embargo period, and

4. the petabytes of the raw LHC data are restricted and preserved for the long-term.

CERN Analysis Preservation tries to encapsulate the policies presented above into a single, easy-to use platform, that can be used for all the steps that scientists need to take in order to preserve and reuse their analysis components. The preservation of raw data is not the domain of CAP, as the storage capacities required exceed 1 petabyte per day.

2.2 Integration Features

2.2.1 Integrations with External Services

As mentioned in the introduction, an important target for CERN Analysis Preservation is to provide an end-to-end tool for scientific preservation, which includes not only the experimental code and datasets, but also a variety of metadata, like researcher affiliations or provenance information. In order to accommodate researchers, CAP provides a variety of integrations with external services, including:

- **Zenodo** [6] is a multidisciplinary open source platform developed and hosted at CERN. The integration with Zenodo makes it possible to generate **Digital Object Identifiers (DOIs)** for software through the CAP submission form by fetching the code from GitLab, for example, and then uploading it to Zenodo where a DOI can be registered.
- The **CERN Document Server (CDS)** [7] and **Indico** [8], the internal CERN services for media and file storage, and appointment scheduling.
- The **Research Organization Registry (ROR)** [9], is a service that provides institution details, which helps unifying affiliations in publications and conferences.
- **ORCID** [10] is a service that provides a persistent digital identifier for each researcher, enabling them to distinguish themselves and keep a persistent record of all their professional information. ORCID's use is widely adopted at CERN [11].

2.2.2 Code Preservation & Git Repositories

Software repositories and version control have been key features of software development in the last decades, and CERN Analysis preservation fully supports their integration. The two most well-known version control platforms, **GitHub** and **GitLab** are providing **webhook integration** [12] through their respective APIs. The users can connect their GitHub/GitLab accounts with the service and link their repositories to an analysis, allowing for instant and automatic updates, on specified repository events, i.e. releases, or pushing new code. This feature is an important time-saver for scientists, as it combines the ease of use of a general and very popular version control system with the ability to run always up-to-date code, if they choose to do so.

2.2.3 Computational Workflows & REANA Reproducible Analysis Platform

Although capturing metadata and code from external services are two necessary steps for describing and preserving a physics analysis, it is also very important to attach the information about the computing environment and how the data processing and data analysis were

performed. One way to achieve this is through containerising (e.g. using Docker images) the analysis steps and preserving the computational workflow steps and their inputs.

Making a research data analysis reproducible basically means to provide structured 'runnable recipes', addressing the following: where is the input data, what software was used to analyse the data, which computing environments were used to run the software, and which computational steps were taken to run the analysis. To achieve this, the **REANA service** was integrated with CAP, to give the possibility to attach the **computational workflows**, the **input** and the **output results** of various steps of an analysis and link them with the metadata in perpetuity.

REANA is a reusable and reproducible research data analysis platform [13]. It helps researchers to structure their input data, analysis code, containerised environments and computational workflows so that the analysis can be instantiated and run on remote compute clouds.

REANA was born to target the use case of particle physics analyses, but is applicable to any scientific discipline. The system paves the way towards reusing and reinterpreting preserved data analyses even several years after the original publication.

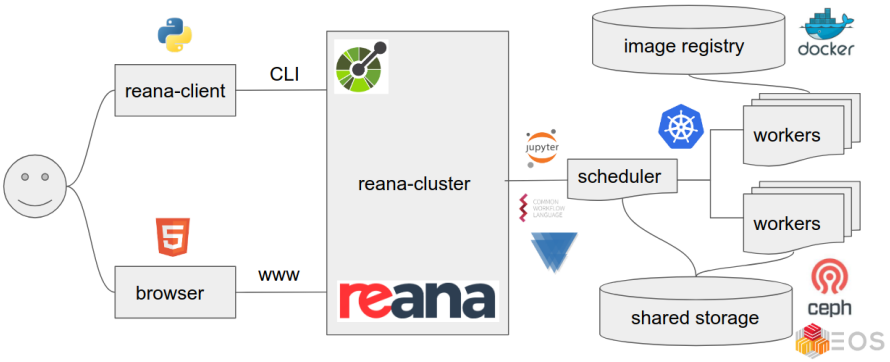

Figure 2: An overview of the REANA platform. The user can interact with the platform via a command line and Web clients. The platform dispatches users' requests to workflow engines and runs necessary tasks on the containerised cloud. The workflow tasks share the same workspace throughout the run.

2.3 FAIR Data

Capturing all the information mentioned above and to achieve better preservation of analyses, the metadata are treated with **FAIR** (Findable, Accessible, Interoperable, Reusable) principles in mind. This means that the data are described by rich metadata and a plurality of relevant attributes, following domain-relevant community standards. Through the experiment-specific schemas and by promoting association of information we capture with detailed provenance, we aim for data conformity and data that are usable and reusable to the collaboration and the researchers forever.

By handling data collections in a FAIR way and by integrating DOIs and other **Persistent Identifiers (PIDs)** from CERN internal and external services, users can connect their analysis components together in one place in a persistent way.

It should be noted that while CAP tries to integrate FAIR and Open Science principles where possible, in order for the service to be useful for our users, control over public sharing

of outputs always remains with the researchers or collaborations, respecting embargoes and other internal procedures (e.g. reviewing).

2.4 Advanced Search Features

The provision of all the above integrated services, as well as experiment-specific databases, allows CERN Analysis Preservation to become a powerful aggregator of most of the information that a researcher could need. This is why all this information is combined in appropriate mappings, and indexed in an **ElasticSearch** cluster, which is the main search engine of CAP.

Triggers, datasets, filters, and other information that was only available through legacy interfaces, is now available through CAP, which can provide fast retrieval of datasets, along with useful dates and statistics, with simple, free-text queries. The ease of use, greatly facilitates the collaboration between theory and practice.

2.5 Technical Details

2.5.1 System Architecture

The service is a web application that follows the pattern of microservice architecture. The implementation language of choice for the backend is **Python** [14], as it offers an assortment of useful data-specific modules and libraries. The platform is built on top of **Invenio** [15], a large-scale digital library framework, that provides an integrated solution for data repositories and integrated library systems, through an ecosystem of compatible modules (based in **Flask** [16] and **SQLAlchemy** [17]).

PostgreSQL [18] is used as CAP's Relational Database Management System. Its main advantage over other databases is the native support for **JSON**, which is the format of choice for modeling the metadata, and encompasses complicated relations and data types. Content and metadata stored into the system are modeled, validated and checked for compliance with community standards by using **JSON Schemas** [19]. This way, everything is then ready to be indexed and become searchable, through our **ElasticSearch** [20] cluster, our search infrastructure that is there to satisfy the information retrieval and recommendation features of the service. All the records are indexed, along with metadata from various other sources, e.g. datasets from external CERN databases.

For the handling of most asynchronous and background tasks, the use of **Celery** [21], a distributed system that acts task scheduler, was introduced, together with the help of **RabbitMQ** [22] as the default message broker. Finally, **Redis** [23] is used as a caching mechanism for the system to handle sessions, permissions, etc. The main web interface is a **ReactJS** [24] single page application, written in **Javascript** [25]. The design follows a responsive paradigm, making it browsable in most browsers and screen sizes.

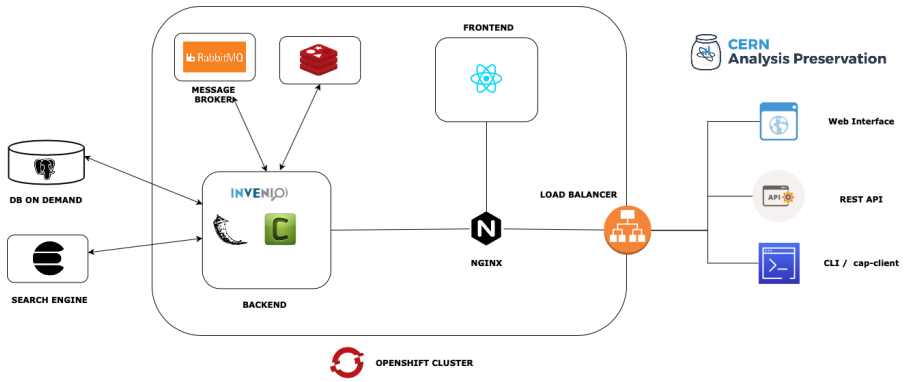

Figure 3: General architecture of CERN Analysis Preservation. The diagram is not exhaustive, as CAP interacts and depends on a variety of external APIs and CERN services.

2.5.2 Personal Data Handling

Metadata and data collected through user submission or automated processes are all retrieved either from open/public databases and repositories, and therefore do not require confidential handling, or from closed/restricted sources, for which data accessibility and retrieval follow the same conditions as those specified at the source. For the use of the service and the classification of users' access rights, standard institutional user authentication is in place, which associates users with their CERN account and their group memberships, without making any information available to third parties.

2.5.3 Scalability and Sustainability Considerations

An important target for CERN Analysis Preservation is to allow for the potential scalability of the service. Although the current schemas and data models are built under the specifications of the LHC experiments (ALICE, ATLAS, CMS, LHCb), more users, experiments and working groups are expected to use the service as it exits the beta phase. To accommodate this, CAP is deployed through the Openshift platform provided by CERN, a service that handles the orchestration of Docker containers, that can be adapted to the services needs.

CAP has been designed to be integrated seamlessly into the way experimental teams work, such that preserving analysis artifacts is not an additional burden on researchers, but rather is facilitated through a service which offers efficiencies to users as they conduct their research. Although this activity remains primarily the domain of experimental teams, for the purposes of sustainability it is considered that the delivery of the service and its continued development/evolution should remain the responsibility of the host lab. Given that the preservation of data and associated analysis artifacts is critical to retaining sufficient knowledge to make meaningful use of data in the future, and potentially beyond the life of an experiment, the sustainability of preservation services and of the preserved assets themselves should be considered an organisational imperative, as a core service in the data management/open data ecosystem of a laboratory. In this way, data and analysis preservation becomes part of the archival record and intellectual legacy of an organisation.

2.6 Interacting with the service

It is possible to interact with CAP in 3 ways: through the web interface, by using our API or the command line client.

2.6.1 Web Interface

The main way of using and accessing the service is through the web interface. The user has the ability to create an analysis workspace, populate it with metadata, upload analysis content (e.g. files, data) and link persistently with external entities, such as code repositories or computational workflows. Users are also able to version analyses and have various options for sharing their work within their collaboration, group, etc. Through the web interface, it is also possible to manage user profiles as well as user tokens for 'remote' authorisation.

2.6.2 RESTful API

The core of the CERN Analysis Preservation platform is its powerful RESTful API. Everything from the SPA, to file upload/download, the webhooks, etc. are handled by it. An OpenAPI v3 specification [26] is also available to help with tests, client integrations, tools, and documentation.

2.6.3 CAP Client

The CAP Client is an alternative to the web interface and the REST API that provides easy access to a subset of the core functionality int. It is an installable Python package, found in **PyPI**.

It can be very useful as a basic functionality cli tool, that can be integrated in local workflows and be part of a researcher's Unix pipeline.

3 Example Use Cases

In general, preservation efforts and needs at CERN change from one experiment to another. Even in the same experiment/collaboration, there are different working groups that work and do analyses in their own way. With CERN Analysis Preservation, we have built a tool that can accommodate these needs, depending on the specifications and requirements individual groups have (e.g. their own data models).

Currently the platform provides content types and use cases for the 4 LHC experiments (ALICE, ATLAS, CMS, LHCb). Each of them, has a customised data model, and is using the platform for their individual needs and their collaboration's internal systems/databases/APIs that are already in place. Some examples of use cases include:

- **CMS**: The CMS experiment's needs are mainly focused on metadata and integration of their experiment databases inside the CAP platform for easier information submission, and better and faster search functionalities [4]. CAP is able to harvest and query their dataset and analysis databases (DAS, CADI), helping connecting experiment entities with the ones inside CAP. CMS are also working on preserving containerised computational workflows through REANA as a stating point.
- **ATLAS**: The use case for ATLAS is mostly centered around preserving containerised computational workflows (through REANA), as well as connecting with their own systems (e.g. Glance) [27].

4 Conclusion

In this paper we presented **CERN Analysis Preservation**, a platform that follows the FAIR principles for data management, and enables the preservation and reuse of research data. CAP is a digital repository platform that enables LHC experiments to store information about analyses and related assets, the user code, the container image, individual n-tuples and plots. CAP integrates with the **REANA** reproducible analysis platform for reinterpreting preserved data and analysis workflows. The CAP and REANA services were designed with ALICE, ATLAS, CMS and LHCb experiment use cases in mind.

The pilot experiments allowed to identify a clear need for custom metadata models targeted to each use case scenario. It is our vision to empower LHC collaborations and researchers to use the CERN Analysis Preservation service as a generic and extensible tool, each community applying their own preservation rules, data models and schema mappings, in order to seamlessly integrate analysis preservation and reproducible science practices and tools into their daily workflows.

References

[1] European Commission, *Open science* (2019), https://ec.europa.eu/research/openscience/index.cfm

[2] European Commission, *Open data: An engine for innovation, growth and transparent governance* (2011), https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

[3] X. Chen, S. Dallmeier-Tiessen, R. Dasler, S. Feger, P. Fokianos, J.B. Gonzalez, H. Hirvonsalo, D. Kousidis, A. Lavasa, S. Mele et al., *Open is not enough* (2018), https://doi.org/10.1038/s41567-018-0342-2

[4] L.L. Iglesias, *The CMS approach to Analysis Preservation* (2019), https://indico.cern.ch/event/773049/contributions/3476181/attachments/1939670/3215866/The_CMS_approach_to_Analysis_Preservation_-_CHEP2019.pdf

[5] J. Cowton, S. Dallmeier-Tiessen, P. Fokianos, L. Rueda, P. Herterich, J. Kunčar, T. Šimko, T. Smith, *Open data and data analysis preservation services for LHC experiments* (2015), https://iopscience.iop.org/article/10.1088/1742-6596/664/3/032030

[6] K. Nowak, L.H. Nielsen, A.T. Ioannidis Pantopikos, *Zenodo, a free and open platform for preserving and sharing research output.* (2016), https://doi.org/10.5281/zenodo.51902

[7] L. Marian, J. Caffaro, J.Y. Le Meur, *Multimedia and Document Services* (2013), https://cds.cern.ch/record/1983613

[8] P. Ferreira, T. Baron, C. Bossy, J. B, M. Pugh, A. Resco, J. Trzaskoma, C. Wachter, *Indico: A Collaboration Hub* (2012), https://doi.org/10.1088/1742-6596/396/6/062006

[9] M. Gould, *Hear us ror! announcing our first prototype and next steps* (2019), https://doi.org/10.5438/cykz-fh60

[10] L.L. Haak, M. Fenner, L. Paglione, E. Pentz, H. Ratner, *ORCID: a system to uniquely identify researchers* (2012), https://doi.org/10.1087/20120404

[11] X. Chen, *ORCID integration at CERN* (2017), https://doi.org/10.5281/zenodo.556917

[12] *GitHub Webhooks [software]*, [Online; accessed 20-02-2020], https://developer.github.com/webhooks/

[13] T. Šimko, L. Heinrich, H. Hirvonsalo, D. Kousidis, D. Rodríguez, *REANA: A system for reusable research data analyses* (2019), https://doi.org/10.1051/epjconf/201921406034

[14] G. Van Rossum, F.L. Drake Jr, *Python tutorial* (1995), https://www.python.org

[15] T. Simko, J. Kuncar, L.H. Nielsen, *Using Invenio for managing and running open data repositories* (2017), https://cds.cern.ch/record/2273317

[16] A. Ronacher, *Flask: A Python Mircoframework [software]* (2010 - 2020), [Online; accessed 10-02-2020], https://palletsprojects.com/p/flask/

[17] M. Bayer, *SQLAlchemy: The Database Toolkit for Python [software]* (2006 - 2020), [Online; accessed 10-02-2020], https://www.sqlalchemy.org/

[18] PostgreSQL Global Development Group, *PostgreSQL [software]* (1996 - 2020), [Online; accessed 10-02-2020], https://www.postgresql.org/

[19] *JSON Schema [software]*, [Online; accessed 07-03-2020], https://json-schema.org/

[20] S. Banon, *ElasticSearch [software]* (2010 - 2020), [Online; accessed 10-02-2020], https://www.elastic.co/products/elasticsearch

[21] *Celery: A distributed task queue [software]* (2009 - 2020), [Online; accessed 19-02-2020], http://www.celeryproject.org/

[22] Pivotal Software, *RabbitMQ [software]* (2007 - 2020), [Online; accessed 19-02-2020], https://www.rabbitmq.com/

[23] S. Sanfilippo, the Redis Labs, *Redis [software]* (2009 - 2020), [Online; accessed 19-02-2020], https://redis.io/

[24] J. Walke, *ReactJS [software]* (2013 - 2020), [Online; accessed 10-02-2020], http://reactjs.org/

[25] *Javascript [software]*, [Online; accessed 13-03-2020], https://www.javascript.com/

[26] *OpenAPI Specification*, [Online; accessed 13-03-2020], https://swagger.io/specification/

[27] D. South, *Analysis Preservation for ATLAS: Conclusions of the Analysis Preservation Panel* (2017), https://indico.cern.ch/event/660145/contributions/2694111/attachments/1512405/2358975/dcc_240817.pdf