Open Data and Data Analysis Preservation Services for LHC Experiments

J Cowton[1,4], S Dallmeier-Tiessen[1], P Fokianos[1,3], L Rueda[1], P Herterich[1,2], J Kunčar[1], T Šimko[1], T Smith[1]

[1] CERN, Switzerland
[2] Humboldt-Universität zu Berlin, Germany
[3] National and Kapodistrian University of Athens, Greece
[4] Northumbria University, United Kingdom

E-mail: tim.smith@cern.ch

Abstract. In this paper we present newly launched services for open data and for long-term preservation and reuse of high-energy-physics data analyses based on the digital library software Invenio. We track the "data continuum" practices through several progressive data analysis phases up to the final publication. The aim is to capture for subsequent generations all digital assets and associated knowledge inherent in the data analysis process, and to make a subset available rapidly to the public. The ultimate goal of the analysis preservation platform is to capture enough information about the processing steps in order to facilitate reproduction of an analysis even many years after its initial publication, permitting to extend the impact of preserved analyses through future revalidation and recasting services. A related "open data" service was launched for the benefit of the general public.

1. Introduction

Research Data Management (RDM) is becoming more and more important in a wide range of scientific disciplines, to ensure long-term preservation and accessibility to data after research projects have ended. Furthermore, making data available is crucial to allow fellow researchers to reproduce scientific results [1]. This movement towards more transparent and open science and scientific communication is increasingly supported by funders who require researchers to submit a data management plan for their projects when applying for a grant [2]. In addition, some publishers ask for the data supplementary to a manuscript to be open in order to consider the paper for publication [3].

These developments are being widely adopted in the High-Energy Physics (HEP) community where all four LHC collaborations have now approved policies on the access and preservation of their data [4, 5, 6, 7]. Referring to the four levels of research data in HEP defined by the Study Group for Data Preservation and Long Term Analysis in High Energy Physics [8], the data policies state that:

- Level 1 data, comprising data that are directly related to publications and provide documentation for the published results, are made available Open Access through suitable community platforms;
- Level 2 data, including simplified data formats, are made available as example analyses for outreach and training exercises;

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

- Level 3 data, covering reconstructed data and simulation data as well as the analysis software needed to allow a full scientific analysis, are made available in a large part after an embargo period of about three to five years;
- Level 4 data, encompassing raw data and providing access to the full potential of the experimental data, are access restricted even within the collaboration and stored for long-term preservation.

While level 1 data have already been shared for several years, a central access point for level 2 and 3 data was still needed to be found.

2. The CERN Open Data Portal

In 2014, the CERN Open Data Portal (home page in Figure 1) was developed as a joint effort between CERN IT and the CERN Scientific Information Service in close collaboration with the experiments.

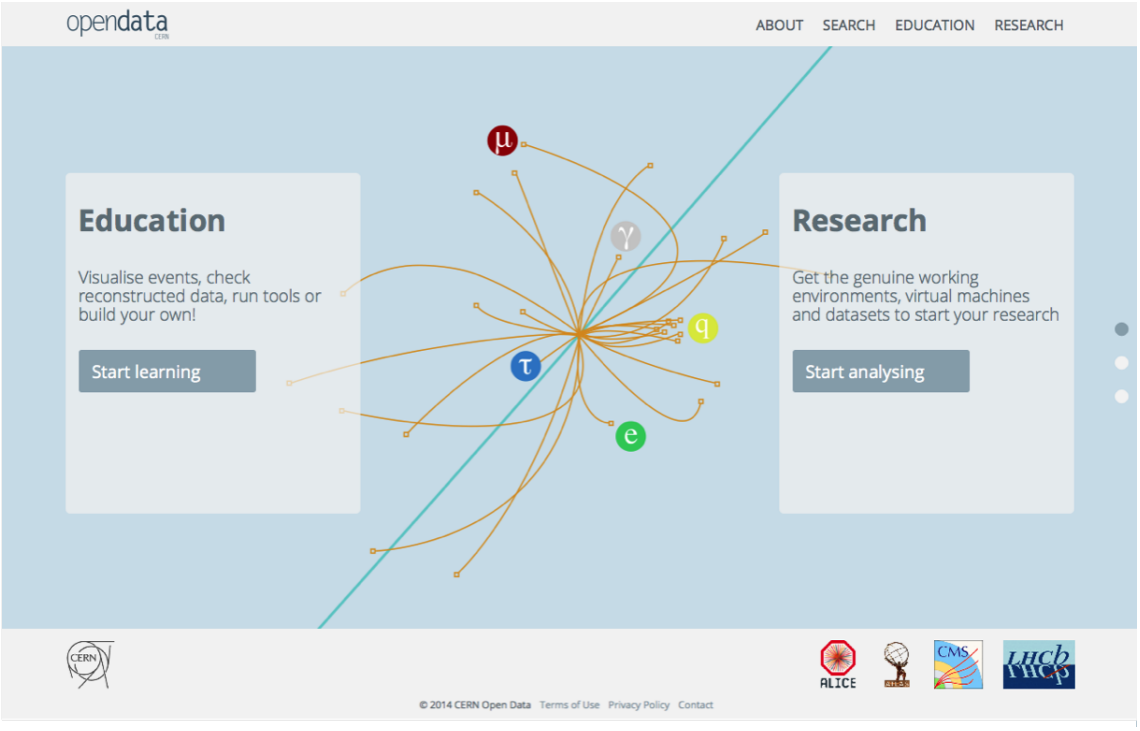

Figure 1. The CERN Open Data Portal Home Page

2.1. Technical infrastructure

The CERN Open Data Portal was built by joining two key building blocks of CERN's information and data services; the digital library software Invenio [9] as the front-end information store and EOS [10] as the back-end file store. Invenio is an open-source software suite which covers all aspects of digital library management, from document ingestion through classification, indexing, and curation up to document dissemination. Invenio complies with standards such as the Open Archives Initiative for metadata exchange and is inspired by practices such as OAIS for information preservation. EOS is a disk-based service providing a low-latency storage infrastructure for physics users. EOS provides a highly-scalable hierarchical namespace, with

data access provided through the XROOT protocol. The main target area for the service are physics data analysis use cases often characterized by many concurrent users, a significant fraction random data access and a large file open rate.

2.2. Content, metadata and usability

The CERN Open Data Portal is now the main access point to level 2 and 3 data from all four LHC collaborations. It currently contains half of the CMS level 3 data, so called primary datasets, taken in 2010 (27 TB of data in total) as well as examples of user analysis code illustrating how the general public could write their own code to perform further analyses on these data. The portal offers several high-level tools which help to visualise and work with the data, such as an interactive event display, shown in Figure 2, permitting to visualise CMS detector events on portal web pages, or a basic histogram plotting interface permitting to create live plots out of CMS reduced datasets [11]. The platform guides high-school teachers and students to online masterclasses to further explore the data and improve their knowledge of particle physics. It also offers the download of Virtual Machine images permitting users to start their own working environment in order to further explore the data; for this the platform uses CernVM [12] based images prepared by the collaborations.

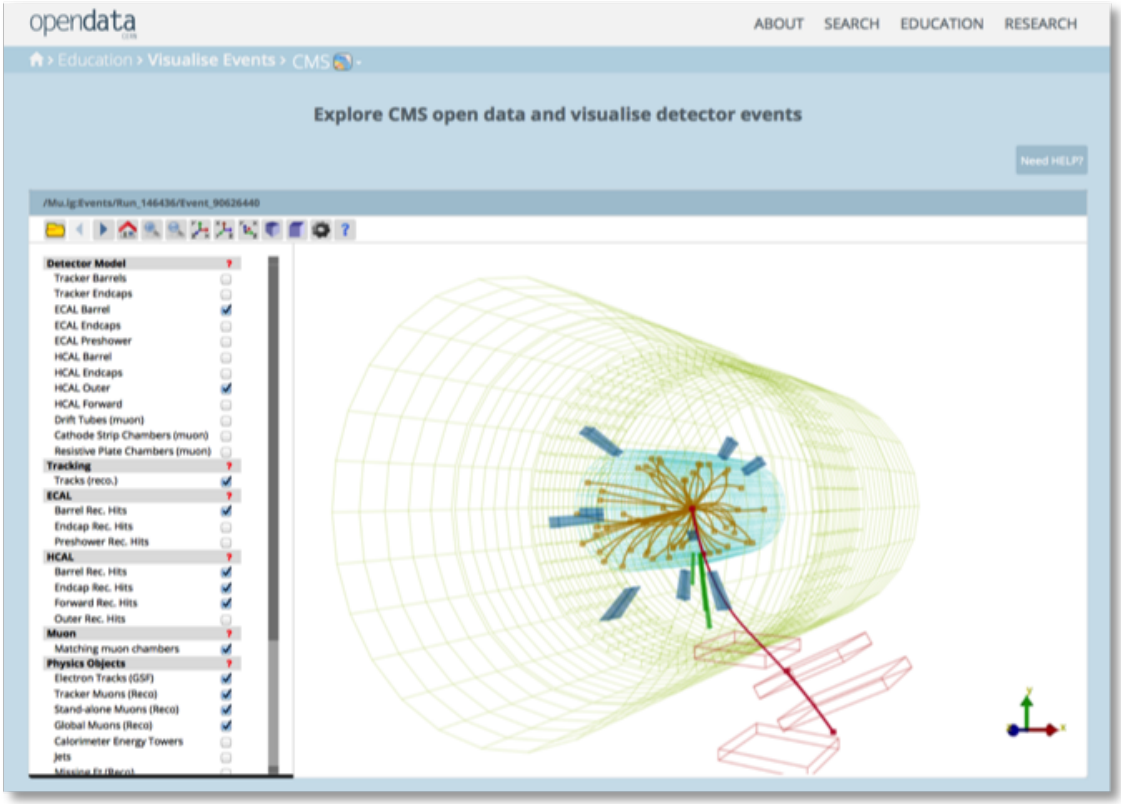

Figure 2. The CMS Interactive Event Display

All the data and tools are accompanied by detailed documentation and rich metadata. The metadata representation of data and software uses the bibliographic standard MARC21 which had to be customised and extended in relation to specific fields for technical metadata or contextualisation. Each entry in the portal is furthermore minted with a Digital Object Identifier

(DOI), ensuring later referencability and citability of preserved data and software according to the FORCE11 Joint Declaration of Data Citation Principles [13].

To provide easy access to the portal's contents a considerable effort was devoted to the design of an effective information architecture and an attractive presentation. The home page, shown in Figure 1, helps the users to select between two profiles: Education, where the visualisation tools and learning resources are the most prominent content; and Research, focusing on providing the environment and tools to kickstart new research projects.

Both entry points were heavily tested and refined to optimise the interface for each profile. Students from the Lapland University of Applied Sciences in Finland as well as groups of young and senior researchers at CERN reviewed the content of the portal, tested the tools and confirmed that all the examples were reproducible. This approach enhanced the quality of the content and the interface.

2.3. Launch and impact

On November 20th 2014, the CERN Open Data Portal was launched officially [14]. In the first month after the launch, the site was visited by 82,000 users of whom 21,000 viewed the data in more detail. Almost 20,000 visitors used one of the tools (event display or histogramming) offered by the portal. On average, the page is used by 1,000 people per day of which 40% look at the detailed data records and 1% download a level 3 data set.

Though none of the datasets is cited with a DOI yet, we know through direct feedback about new use cases of published datasets, such as new collaborations based on the published data, adoption of the provided code examples for new analyses and the (re-)use of the primary datasets for machine learning on big data. If a link to a published result is provided, these (re-)use cases will be included in the portal for further inspiration.

3. CERN Analysis Preservation Framework

Sharing data openly is just the last step in a chain of good data management practices and thus, the CERN Open Data Portal presents the tip of an iceberg of services provided by CERN to the LHC collaborations. To enable full reproducibility of research results, a thorough documentation of the process is needed. In addition, all the tools used for doing the analysis need to be preserved so they can be employed again to reproduce or replicate a result. The CERN Analysis Preservation Framework is developed as the central platform for all four LHC collaborations to preserve information about and tools for analyses.

3.1. Approach

Before starting the development, an analysis preservation pilot study was launched in order to assess the usual workflow practices in LHC collaborations. Leveraging on synergies between ALICE, ATLAS, CMS and LHCb experiments, the analysed data will be captured through various analysis and processing steps, from the initial capture and pre-selection of primary data, through several intermediate selection steps yielding greatly reduced datasets, up to the final selection of N-tuples used for producing high-level plots appearing in scientific journal publications. Most of the analysis chain is kept strictly restricted within a given collaboration; only the final plots, presentations and papers are usually made public. It is therefore essential to handle access rights and embargo periods as part of the data life cycle.

The study revealed many similarities between collaborations, even though the variety of different practices existing in different groups within the collaborations make it hard to reproduce an analysis at a later time in a uniform way. One recurring problem underlined by the study was to ensure an efficient "knowledge capture" related to user code when the principal author of an analysis (e.g. a PhD student) leaves the collaboration later. In addition, there is a common need to be able to connect published results to the analyses activities within the collaborations.

The pilot solution has been prototyped using the Invenio digital library platform which was extended with several data-handling capabilities. The aim is to preserve information about datasets, the underlying OS platform and the user analysis code used to study and process it. The configuration parameters, the high-level physics information such as physics object selection, and any necessary documentation and discussions are optionally being recorded alongside the process as well. Thus, the analysis preservation framework will be able to capture enough information about the process in order to facilitate reproduction of an analysis even many years after its initial publication, permitting to extend the impact of preserved analyses through future revalidation and recasting services.

3.2. Challenges for the framework

The analysis preservation framework should be easy to use and to integrate in the every day research workflow. An analysis will be entered through a submission form. To make this process as smooth as possible, the framework will be connected to several collaboration databases that will auto-fill as much information as possible. The submission form follows the logic and steps of the analysis and is tailored to the respective specifics of each collaboration. An example subsection of a submission form is shown in Figure 3.

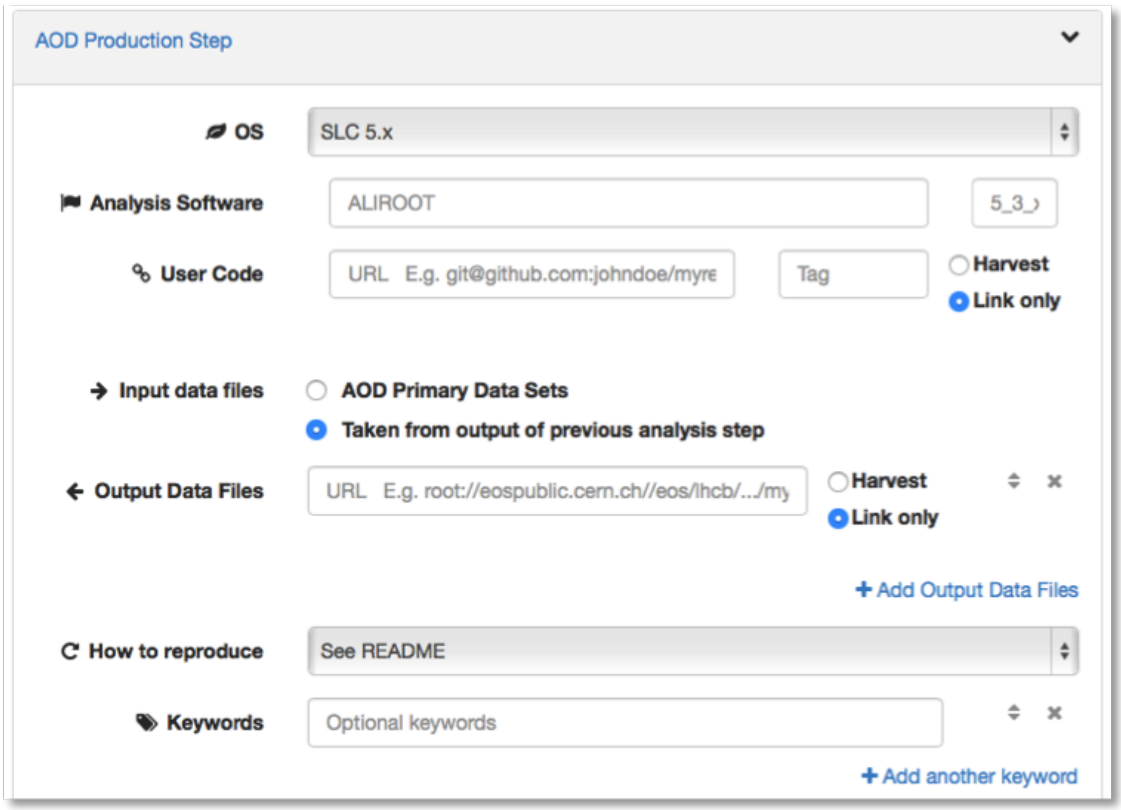

Figure 3. Example of Customised Metadata Capture

Unlike for the CERN Open Data Portal, the metadata schema for the analysis preservation framework will use JSON to accommodate the complex metadata in the best way possible. JSON is commonly used, also by many of the collaboration databases and can be easily extended to JSON-LD to support Linked Data. The data models developed for the analysis preservation

framework will be the starting point of a collaboration with DASPOS [15] on the development of an ontology modelling data analysis and preservation in High-Energy Physics.

3.3. Challenges for the technology

The main developments are to establish interoperability with a variety of data and information sources, based on an assortment of different technologies, with and without remotely usable APIs. Connectors are being built in the platform to large-scale data storage systems (EOS, CASTOR, Ceph), and schemes for streaming data in the front-end or delegating streaming to the back-end are being established. In addition, connectors are being built to the internal information management systems of LHC experiments (e.g. CMS CADI), to the discussion platforms (e.g. TWiki, SharePoint), and to the final publication servers (e.g. CDS, INSPIRE) used in the process.

3.4. Next steps

A first fully functioning version of the framework is expected to be available for testing in summer 2015. That version will be tested by the collaborations to make sure that all information needed to reproduce an analysis is captured. Based on the feedback of these tests, the usability of the submission forms will be improved and every attempt made to automate the capture of information rather than requiring extra manual steps from the physicists. This will inevitably mean the construction of more and more connectors to additional information sources. The service will be deployed in a scalable architecture of load-balanced front-end and back-ends and will be scaled to match the growing usage.

4. Outlook

For analysis capture to become a standard part of the scientific process we aim to integrate the framework into collaboration workflows so that information can be captured seamlessly at as early a stage as possible. We will be working with individual experiments to encourage adoption by integrating checks of the captured information into approval processes for future publications. We will also be investigating how best to capture information about completed analyses before the data, tools or knowledge is lost as authors leave the collaborations. This will also provide a showcase of available analyses and tools which can prove useful within the collaborations as training resources for new students, and to raise awareness of the importance of preserving analyses for future generations.

Acknowledgments

Building these portals has been a truly collaborative effort of many people from many different teams, and we would like to thank all our collaborators. In particular:
from CMS: Kati Lassila-Perini, Tom McCauley, Achintya Rao, Alicia Calderon, Ana Rodriguez-Marrero, Adam Huffman, Jonatan Piedra
from LHCb: Silvia Amerio, Ben Couturier, Ana Trisovic
from ALICE: Mihaela Gheata, Costin Grigoras
from ATLAS: Kyle Cranmer, Felix Socher, David Rousseau, Lukas Heinrich
from DASPOS: Mike Hildreth, Charles F. Vardeman, Natalie Meyers
from DPHEP: Frank Berghaus, Jamie Shiers
from CernVM: Jakob Blomer
from CERN EOS: Luca Mascetti
as well as Peter Igo-Kemenes, Robin Colignon, and the Lapland University of Applied Sciences

References

[1] Stodden, Victoria, Borwein, Jonathan, and Bailey, David H. (2013). "Setting the default to reproducible". In: computational science research. SIAM News 46, pp. 4-6
[2] National Science Foundation (n.d.). NSF Data Management Plan Requirements. Available at `http://www.nsf.gov/eng/general/dmp.jsp`
[3] Bloom, Theodora, Ganley, Emma and Winker, Margaret (2014). "Data access for the Open Access literature: PLOS's data policy." PLoS medicine 11.2: e1001607. `doi:10.1371/journal.pmed.1001607`
[4] ATLAS collaboration (2014). ATLAS Data Access Policy. CERN Open Data Portal. DOI: `10.7483/OPENDATA.ATLAS.T9YR.Y7MZ`
[5] ALICE collaboration (2013). ALICE data preservation strategy. CERN Open Data Portal. DOI: `10.7483/OPENDATA.ALICE.54NE.X2EA`
[6] CMS collaboration (2012). CMS data preservation, re-use and open access policy. CERN Open Data Portal. DOI: `10.7483/OPENDATA.CMS.UDBF.JKR9`
[7] LHCb collaboration (2013). LHCb External Data Access Policy. CERN Open Data Portal. DOI: `10.7483/OPENDATA.LHCb.HKJW.TWSZ`
[8] DPHEP Study Group (2009). "Data Preservation in High Energy Physics." arXiv preprint. arXiv:0912.0255
[9] Šimko, Tibor *et al*, `http://invenio-software.org`
[10] Peters, Andreas *et al*, available at `http://eos.web.cern.ch`
[11] Mc Cauley, Thomas (2015). "Open access to high-level data and analysis tools in the CMS experiment at the LHC." See contribution in this proceedings
[12] Buncic, Predrag *et al* (2011)."CernVM: Minimal maintenance approach to the virtualization." J. Phys.: Conf. Ser. 331 052004
[13] Data Citation Synthesis Group (2014). "Joint Declaration of Data Citation Principles." Martone M. (ed.) San Diego CA: FORCE11 `https://www.force11.org/datacitation`.
[14] CERN (2014). "CERN makes public first data of LHC experiments." Press Release available at `http://home.web.cern.ch/about/updates/2014/11/cern-makes-public-first-data-lhc-experiments`
[15] Hildreth, Mike *et al*, `http://daspos.crc.nd.edu`