Use case: CERN Analysis Preservation

Sünje Dallmeier-Tiessen[1], Artemis Lavasa[1], Tibor Šimko[1], Javier Delgado Fernández[1,2], Pamfilos Fokianos[1], Robin Dasler[1], Anxhela Dani[1,3], Annemarie Mattmann[1,4], Ioannis Tsanaktsidis[1], Anna Trzcinska[1], Diego Rodriguez Rodriguez[1,2]

***Abstract*— The CERN Analysis Preservation Framework is a central platform for the four LHC collaborations at CERN and it was developed to address the need for the long-term preservation of all the digital assets and associated knowledge in the data analysis process, in order to enable future reproducibility of research results. As the service continues to develop, new challenges arise, so in this paper we present the initial considerations for implementing a common data markup schema for all the information in the service, which is essential for enhanced search and discovery, and for supporting links with various other internal and external platforms.**

LHC data is unique, precious and complex. Consequently, preserving this data, the software and the knowledge around a physics analysis in a comprehensive reusable manner is challenging. The CERN Analysis Preservation (CAP) service was started in partnership with the LHC collaborations (i.e. ALICE, ATLAS, CMS and LHCb) to try and overcome those barriers. It aims to enable individual physicists to preserve information as it is produced by making it easier to describe, preserve, find, exchange and hand over information in a fast paced environment of highly fluctuating personnel.

The CERN Analysis Preservation service (Figure 1) is a response to two parallel demands, internal and external. Within the Collaborations, the high and complex output of analyses and analysers results in significant challenges in terms of capturing and preserving the analysis and the knowledge around it, which is of particular relevance for future reuse and reproducibility of the research results. Externally, an increasing number of funding agencies have put in place data management policies, which demand the development of comprehensive data management frameworks for data and knowledge preservation, and for future reuse or even reproducibility of research outcomes. A similar development can be observed

[1] CERN, Geneva, Switzerland
[2] Universidad de Oviedo, Oviedo, Spain
[3] Alexander Technological Educational Institute of Thessaloniki, Thessaloniki, Greece
[4] Technische Universität Darmstadt, Darmstadt, Germany

among journals, which start making data and reproducibility measures mandatory for any paper to be published.

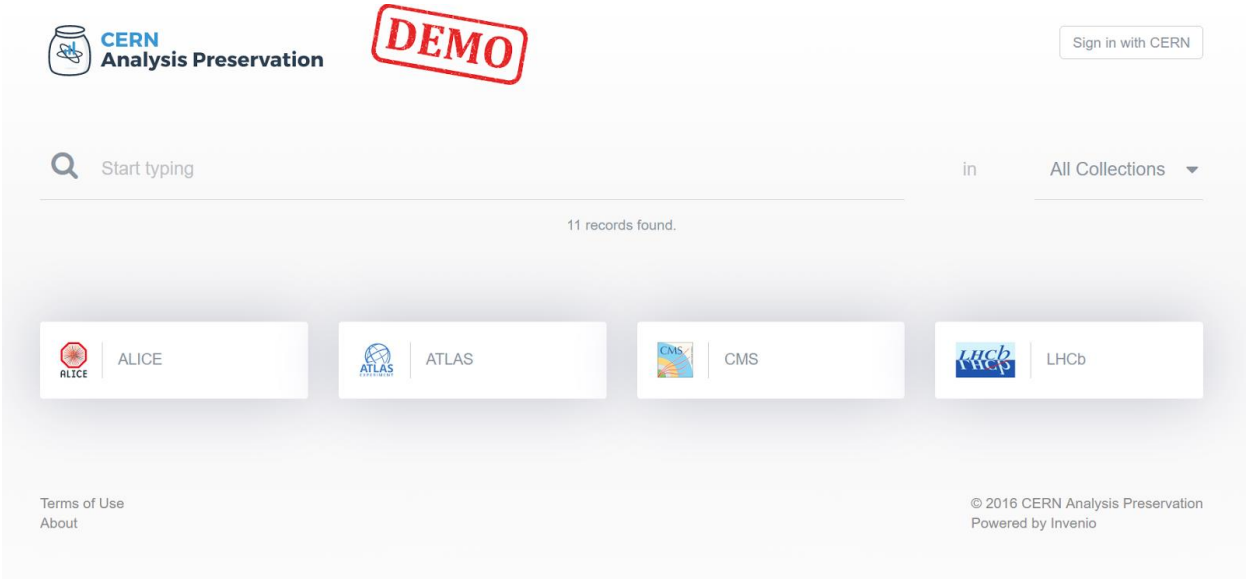

Figure 1: The CAP service

CAP is developed as an open source tool and built on the Invenio Open Source Digital library framework[5], which also underpins services such as Zenodo[6] or CERN Open Data[7] and the data model is developed in JSON[8], which is able to handle highly intricate metadata structures. CAP service functions are constantly being enriched and improved through CERN's participation in the THOR project (Technical and Human Infrastructure for Open Research)[9], a European Commission-funded project under the H2020 framework program; for example, through this project work is being done to implement ORCID authentication and to explore dynamic data citation.

As the tool gathers information throughout the research lifecycle, it is faced with complex materials, such as data, software, their dependencies and versions. To facilitate future reuse of such research objects, it is indispensable to enable the researcher to provide enough context information around the analysis (progress). A standard analysis "record" within the service contains detailed information about the processing steps, the datasets that are used, the software (version) used and, in addition, detailed physics information is made available (e.g. energy level, vetoes, cuts)[10]. It is expected that researchers/users of the

[5] http://invenio-software.org/
[6] https://zenodo.org/
[7] http://opendata.cern.ch/
[8] http://www.json.org/
[9] https://project-thor.eu/
[10] The schemas used: https://github.com/cernanalysispreservation/analysis-preservation.cern.ch/tree/master/cap/jsonschemas

tool will also use this information to search for interesting ongoing analysis (content).

Of course, to be able to benefit from the rich knowledge that could be made available in the tool, it would be best to standardise the information inside as much as possible. This would enable an unprecedented search functionality across these preserved content. Figure 2 below shows an example excerpt used by the CMS collaboration. Currently, the information provided in analysis records is tailored to each experiment, in order to address their particular needs, which makes standardisation even more challenging.

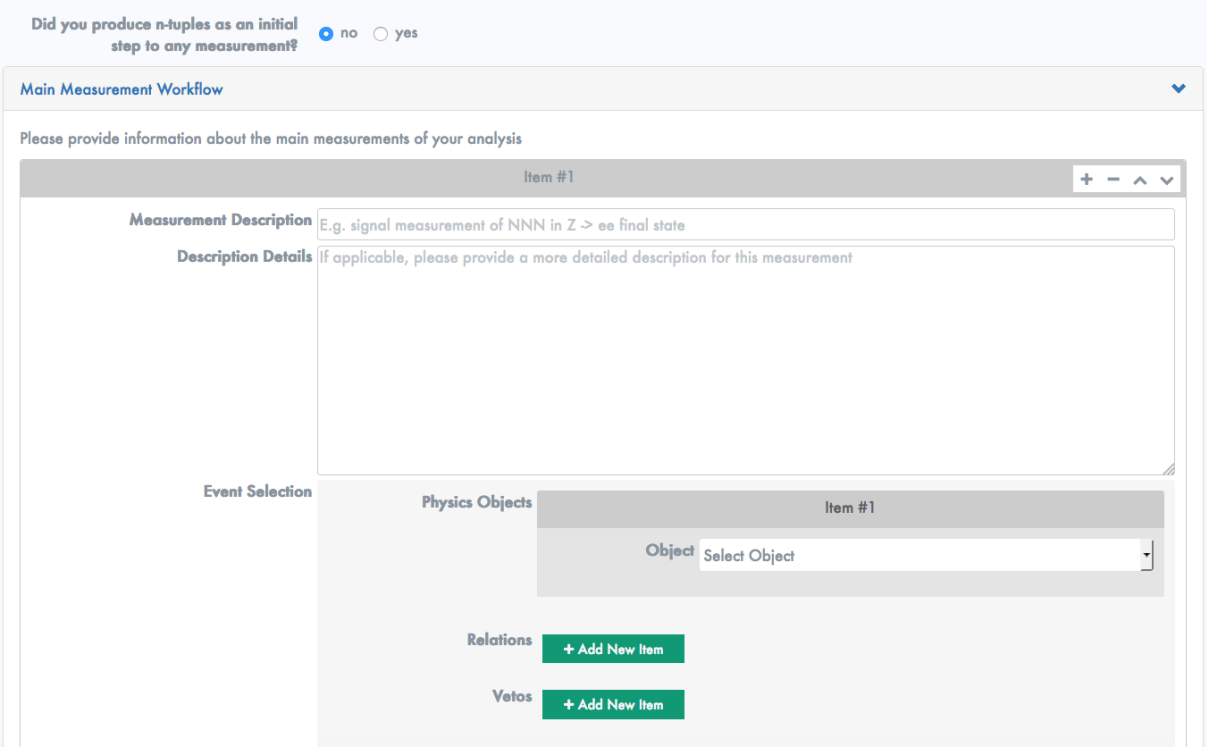

Figure 2: Example excerpt from the schema developed for the CMS collaboration

First attempts have been made to look into existing vocabularies and in particular into using the Schema.org[11] vocabulary encoded in JSON-LD[12], which would provide much-needed support for Linked Data. From this initial research, it became apparent that a vocabulary that could provide wide enough coverage for our purposes without having to be extended significantly to reflect the community specific subject has not been developed yet. This is, of course, often the case when attempting uniformity within any given system with unique characteristics. More specifically, there are terms that would need to be standardised that may only find application in one experiment of the LHC. So, the most logical way to handle this situation would be to integrate a common data vocabulary for the more high-level information

[11] https://schema.org/
[12] http://json-ld.org/

across the platform and then proceed with extensions where needed, most likely on an experiment-by-experiment basis.

As the service continues to grow, more and more records will be introduced into the system, thus making the implementation of a common way of describing all this information more relevant than ever, not only for CERN Analysis Preservation, but for CERN's other platforms as well (e.g. the CERN Open Data Portal). This is particularly crucial to facilitate a meaningful search. As an example of a searchable field that would be facilitated by a common vocabulary and format is the particle reactions and observables for each data table (Figure 3). This has already been implemented in HEPData[13], the High-Energy Physics open-access data repository, which is hosted by CERN.

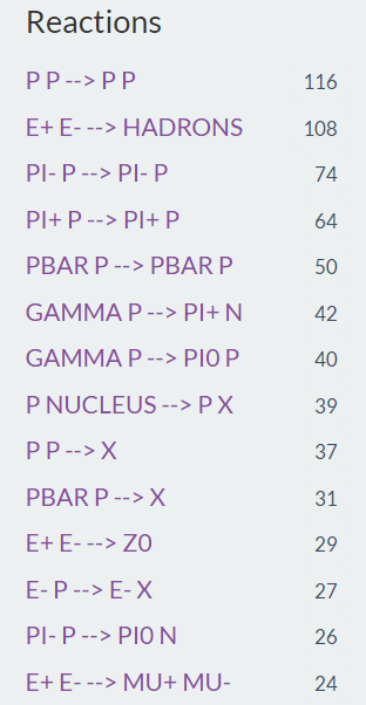

Figure 3: HEPData reactions facet

Some work on this subject to address the need for formally representing and describing experimental results within the particle physics community has already been carried out (Carral et al., 2015). As we work to develop high-quality services for our users, it is essential that we build upon and expand these ideas, and the improvement and semantic harmonisation of our data models is an integral part of that.

[13] https://hepdata.net/

REFERENCES

Carral, D., Cheatham, M., Dallmeier-Tiessen, S., Herterich, P., Hildreth, M. D., Hitzler, P., ... & Watts, G. (2015, October). An Ontology Design Pattern for Particle Physics Analysis. In WOP.