Contribution ID: **466** Type: **Oral**

Invenio digital library technology for open research data repositories

Monday, 10 October 2016 12:00 (15 minutes)

We present the new Invenio 3 digital library framework and demonstrate its application in the field of open research data repositories. We notably look at how the Invenio technology has been applied in two research data services: (1) the CERN Open Data portal that provides access to the approved open datasets and software of the ALICE, ATLAS, CMS and LHCb collaborations; (2) the Zenodo service that offers an open research data archiving solution to world-wide scientific communities in any research discipline.

Invenio digital library framework is composed of more than sixty independently developed packages on top of the Flask web development environment. The packages share a set of common patterns and communicate together via well-established APIs. The packages come with extensive test suite and example applications and use Travis continuous integration practices to ensure quality. The packages are often developed by independent teams with special focus on topical use cases (e.g. library circulation, multimedia, research data). The separation of packages in the Invenio 3 ecosystem enables their independent development, maintenance and rapid release cycle. This also allows the prospective digital repository managers who are interested in deploying an Invenio solution at their institutions to cherry-pick the individual modules of interest with the aim of building a customised digital repository solution targeting their particular needs and use cases.

We discuss the application of the Invenio package ecosystem in the research data repository problem domain. We present how a researcher can easily archive their data files as well as their analysis software code or their Jupyter notebooks via GitHub <-> Zenodo integration. The archived data and software is minted with persistent identifiers to ensure their further citeability. We present how the JSON Schema technology is used to define the data model describing all the data managed by the repository. The conformance to versioned JSON schemas ensure the coherence of metadata structure across the managed assets. The data is further indexed using Elasticsearch for the information retrieval needs. We describe the role of the CERN EOS system used as the underlying data storage via a Pythonic XRootD based protocol. Finally, we discuss the role of virtual environments (CernVM) and container-based solutions (Docker) used with the aim of reproducing the archived research data and analyses software even many years after their publication.

Tertiary Keyword (Optional)

Experience/plans from outside experimental HEP/NP

Secondary Keyword (Optional)

Preservation of analysis and data

Primary Keyword (Mandatory)

Collaborative tools

Primary authors: KUNCAR, Jiri (CERN); NIELSEN, Lars Holm (CERN); SIMKO, Tibor (CERN)

Presenter: NIELSEN, Lars Holm (CERN)

Session Classification: Track 8: Security, Policy and Outreach

Track Classification: Track 8: Security, Policy and Outreach