Erweitertes Forschungsdatenmanagement: Langzeitarchivierung von Forschungsdaten in der Hochenergiephysik

Patricia Herterich
@pherterich
Humboldt-Universität zu Berlin & CERN Scientific Information Service
6. Bibliothekskongress Leipzig 2016
14.03.2016

Vision: Open & Reproducible Research

- Open Research or Open Science aims at making research as transparent and reproducible as possible
  - Open Access to publications
  - Access to data produced or used
  - Access to code and methodology applied
  - Use of Open Source software instead of proprietary solutions whenever possible

Policy framework - CERN

The CERN Convention (1953) contains what is effectively an early Open Research manifesto:

> *"… the results of its experimental and theoretical work shall be published or otherwise made generally available"*

Policy framework – LHC Collaborations

Approved CB 20th June 2014

ATLAS Data Access Policy

May 21st 2014

Introduction

ATLAS has fully supported the principle of open access in its publication policy. This document outlines the policy of ATLAS as regards open access to data at different levels as described in the DPHEP [1] model. The main objective is to make the data available in a usable way to people external to the ATLAS collaboration.

The ATLAS policy for data preservation is described in a separate document. The collaboration's need to preserve data for its own use shares some requirements with making them open access. To support open access to data additional resources will be required to develop and support the tools to make the data available.

Policies fo

Open access to
levels of increa

ur
is

LHCb External Data Access Policy

LHCb Public Note

Issue: 1
Revision: 1

Reference: LHCb-PUB-2013-003
Created: 22th April 2013
Last modified: 22th April 2013

ALICE data preservation strategy

Sunday, October 6, 2013

The data harvested by the ALICE Experiment up to now and to be harvested in the future constitute the return of investment in human and financial resources by the international community. These data embed unique scientific information for the in depth understanding of the profound nature and origin of matter. Because of their uniqueness, long term preservation must be an essential objective of the data processing framework and will lay the foundations of the ALICE Collaboration legacy to the scientific community as well as to the general public. These considerations call for a detailed assessment of the ALICE data preservation strategy and policy. Documentation, long term preservation at various levels of abstraction, data access and analysis policy and software availability constitute the key elements of such a data preservation strategy allowing future collaborators, the wider scientific community and the general public to analyze data for educational purpose and for eventual reassessment of the published results. The present document describes the basic principles that will guide the redaction addressed by the ALICE data preservation policy.

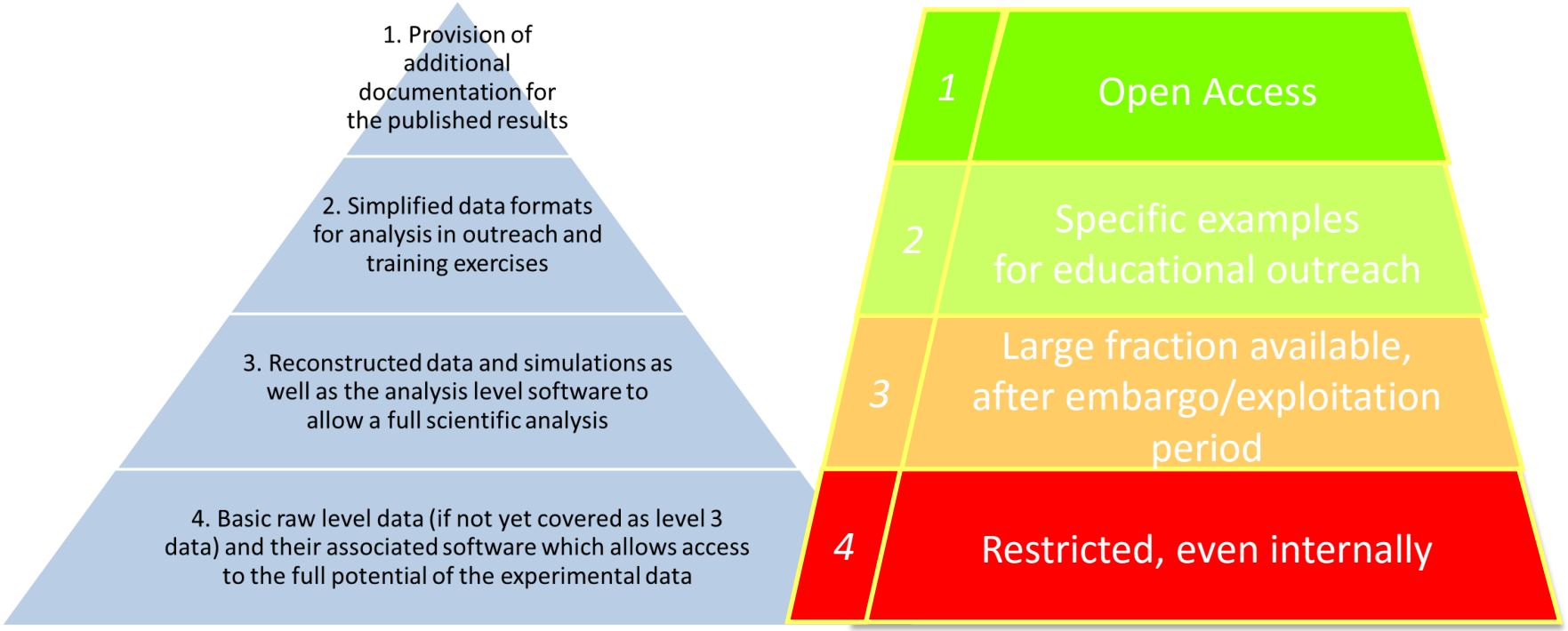
CMS data preservation, re-use and open access policy

CMS data are unique and are the result of vast and long-term moral, human and financial investment by the international community. There is unique scientific opportunity in re-using these data, at different level of abstraction and at different points in time[1]. This opportunity calls for our collective responsibility, and poses unprecedented challenges as no data sample of this complexity and value has ever been preserved or made available for later re-use.

The CMS collaboration is committed to preserve its data, at different levels of complexity, and to allow their re-use by a wide community including: collaboration members long after the data are taken, experimental and theoretical HEP scientists who were not members of the collaboration, educational and outreach initiatives, and citizen scientists in the general public.

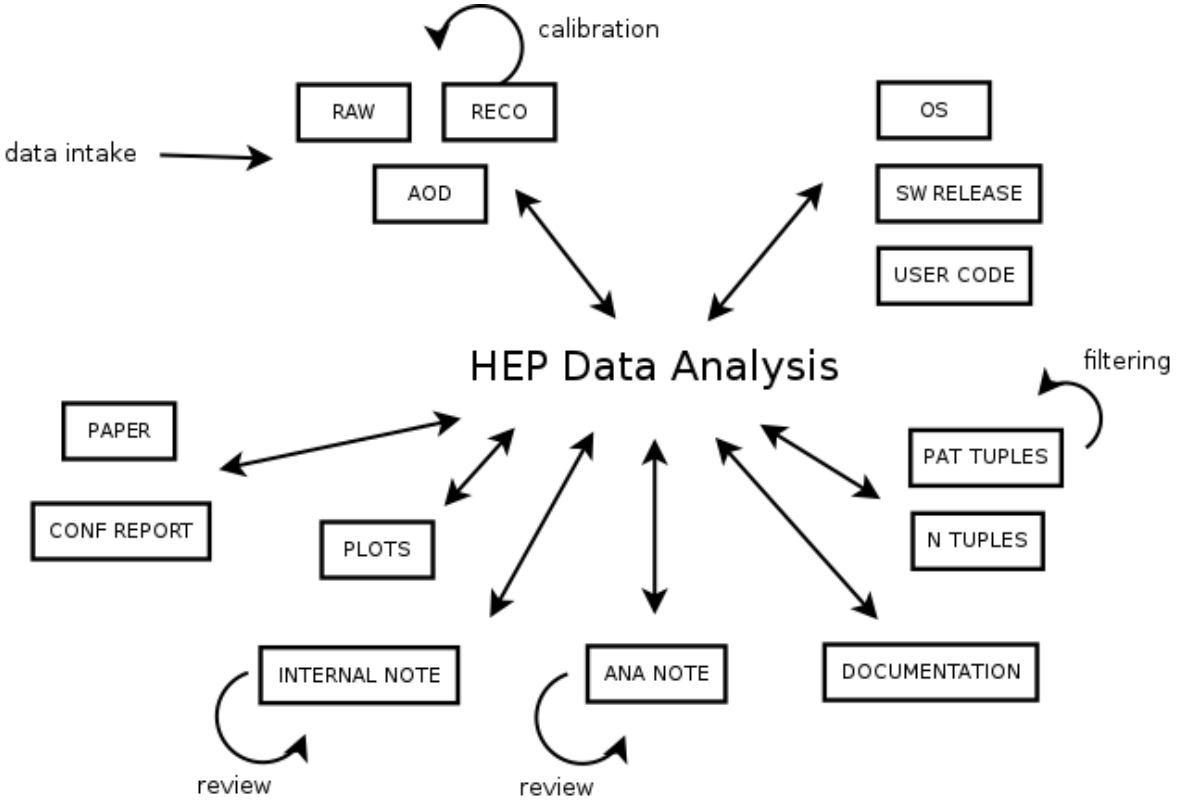
CMS upholds the principle that open access to the data will, in the long term, allow the maximum realization of their scientific potential. To that extent, CMS will provide open access to its data after a suitable but relatively short embargo period, allowing CMS collaborators to fully exploit their scientific potential.

Policy framework – LHC Collaborations

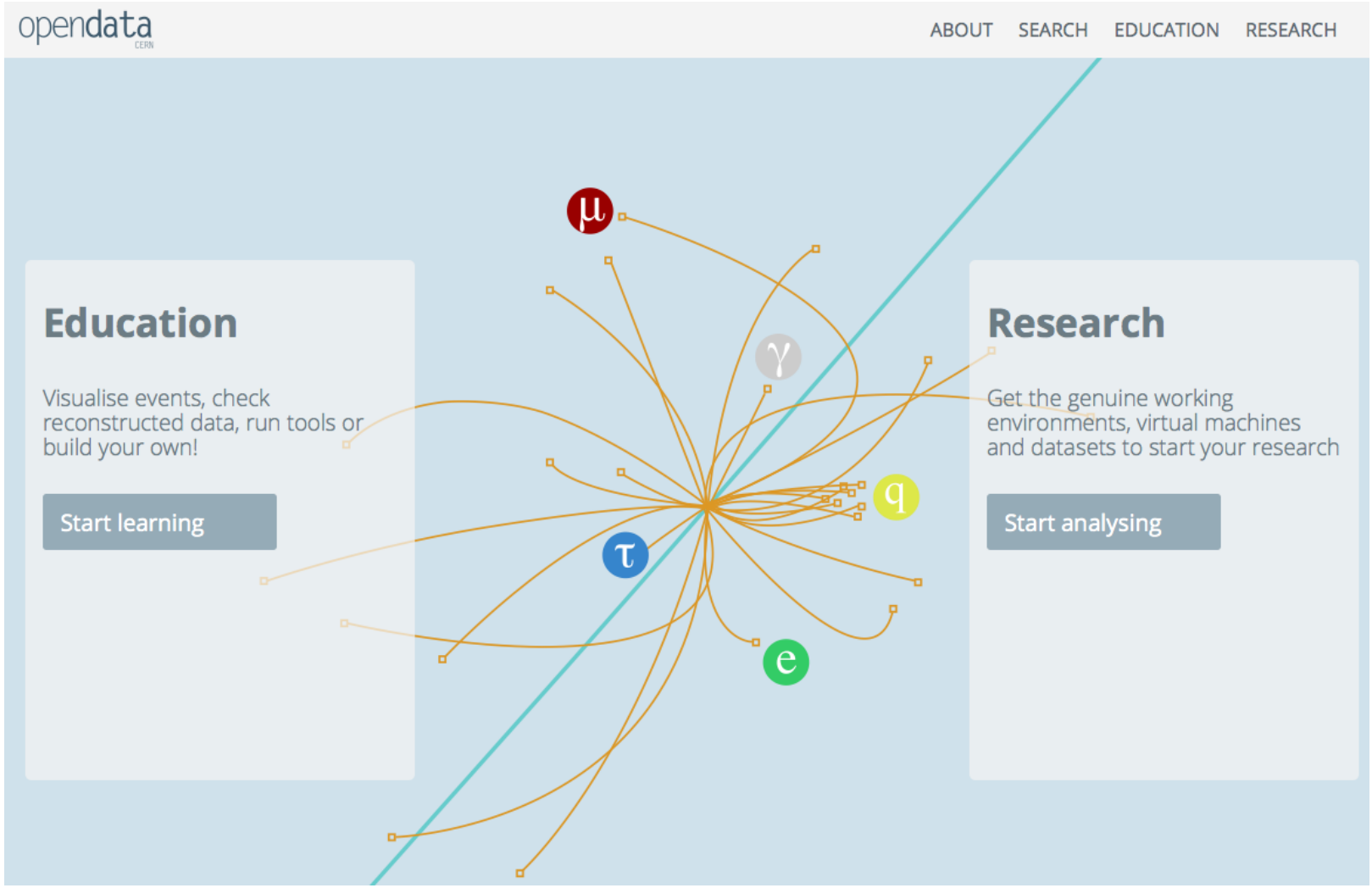


More than research data

"reconstructed data and simulations **as well as the analysis level software to allow a full scientific analysis**"



CERN Open Data Portal

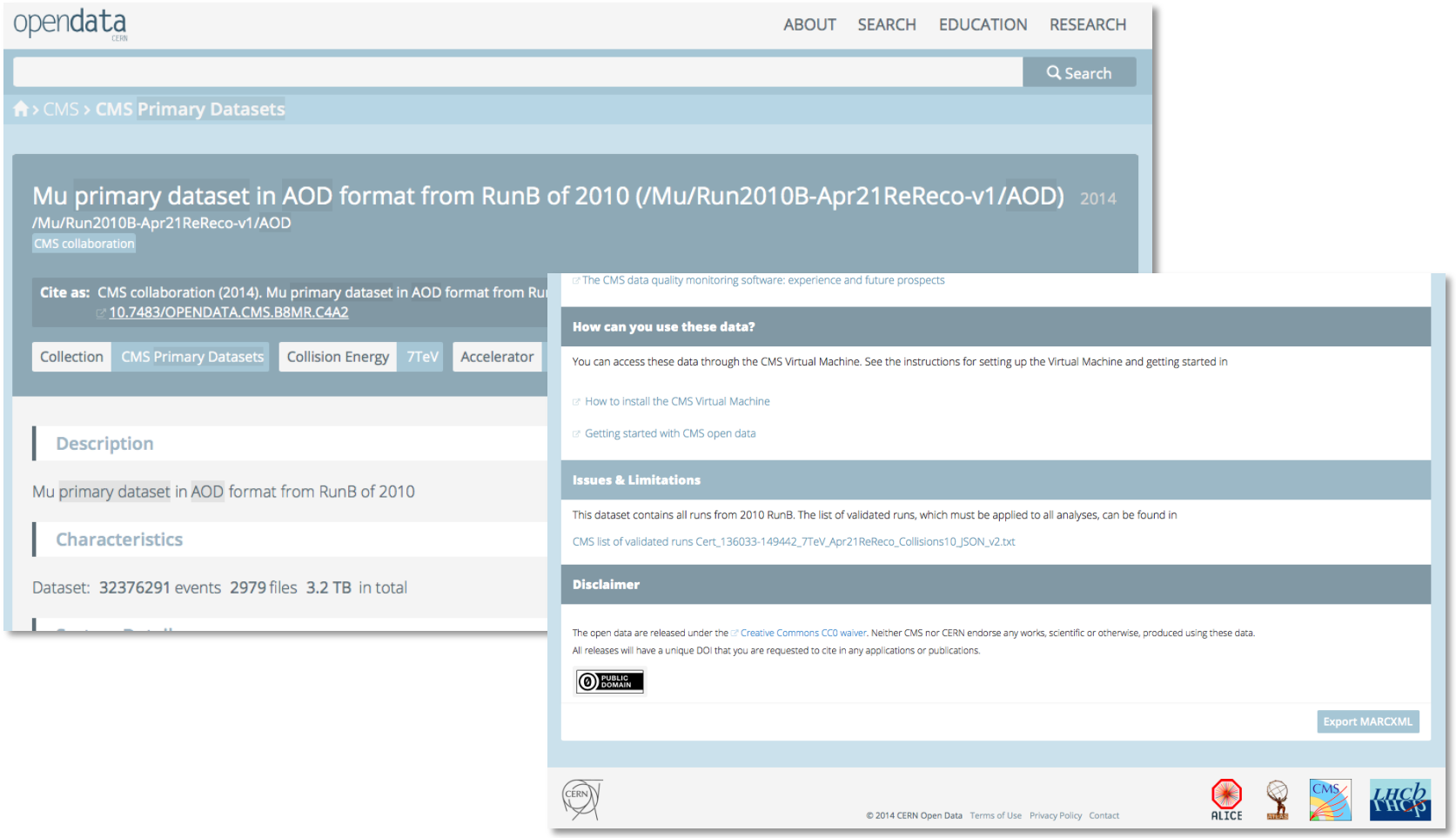


CERN Open Data Portal

- Launched Nov 2014: http://opendata.cern.ch/
- Public access point to data (incl. software and documentation) produced at CERN
- Upon release access to 27 TB of CMS data + educational data from all 4 LHC experiments
- Github repo: https://github.com/cernopendata/
- Based on Digital Library Software Invenio 2.0
- “preservation in the open”

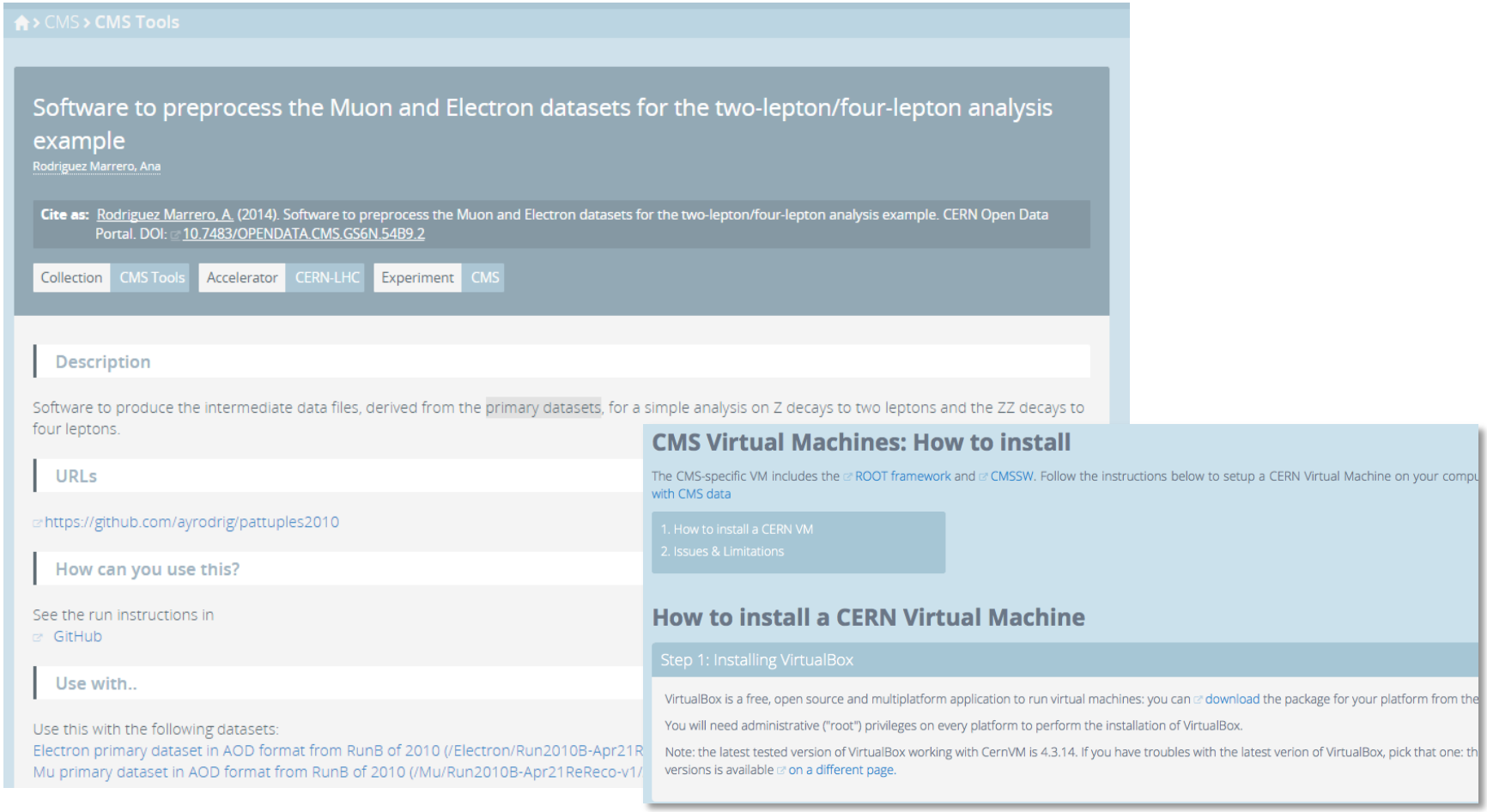
opendata CERN

Data



opendata CERN

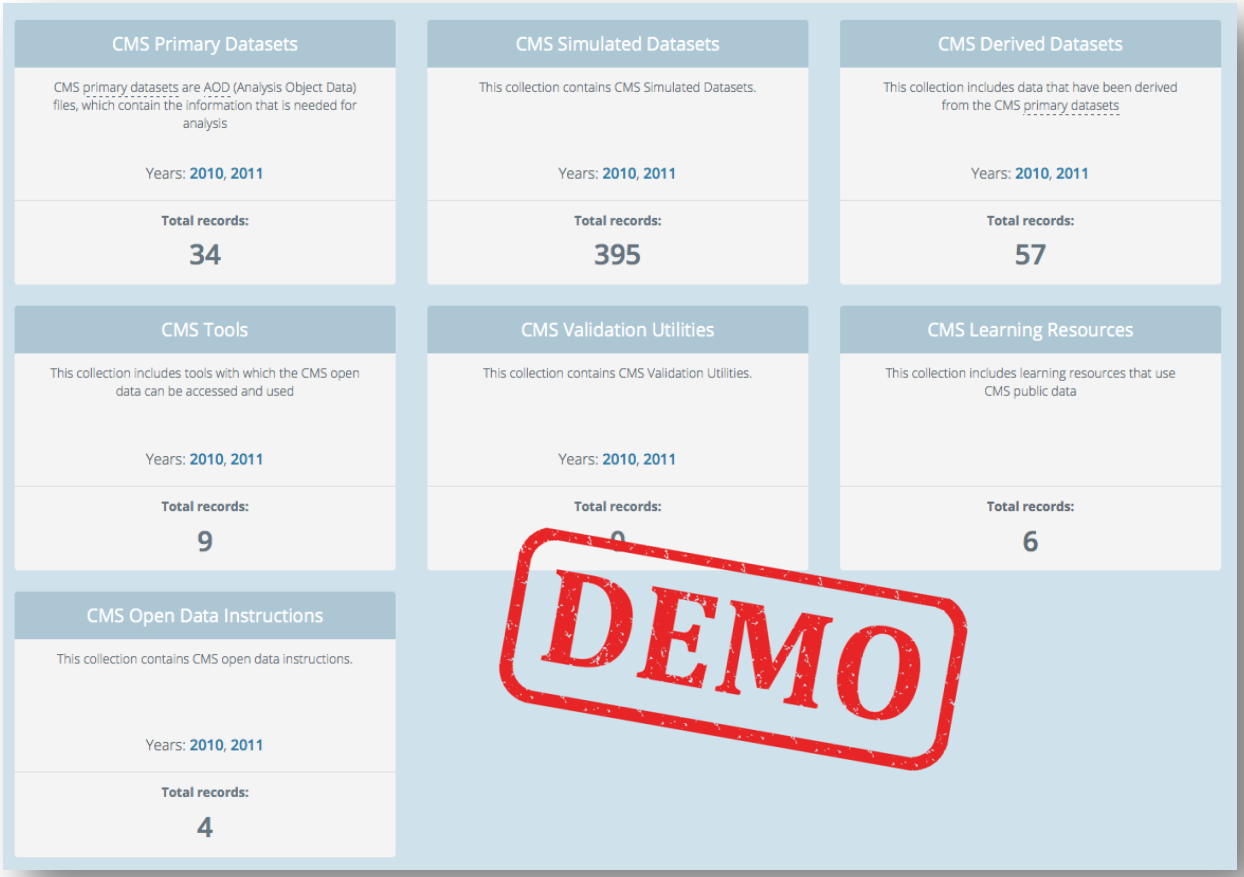
Software & documentation



opendata CERN

Forthcoming: new data releases

270 TB primary and simulated data



opendata CERN

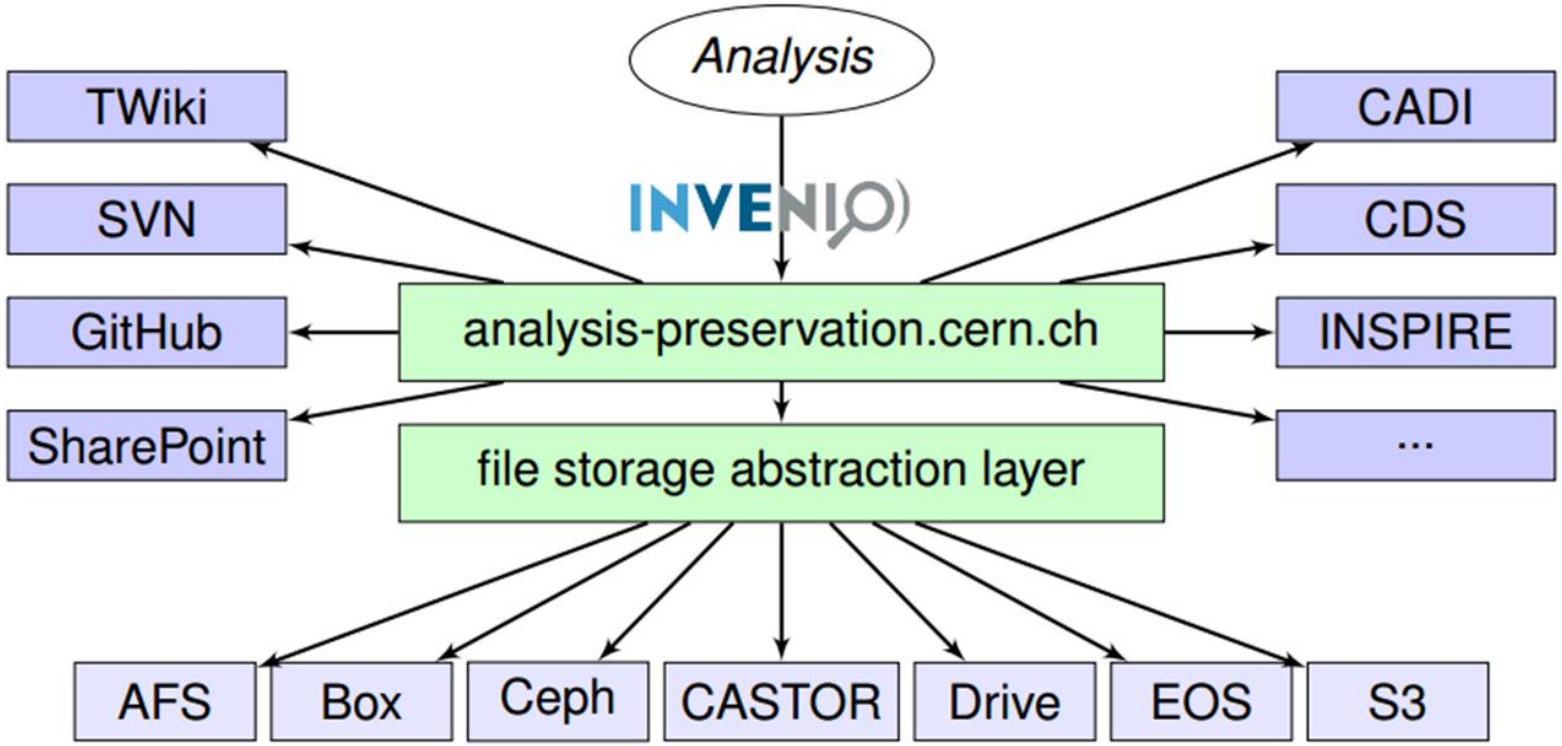
CERN Analysis Preservation

"**closed** counterpart" to CERN Open Data that captures the complexity of

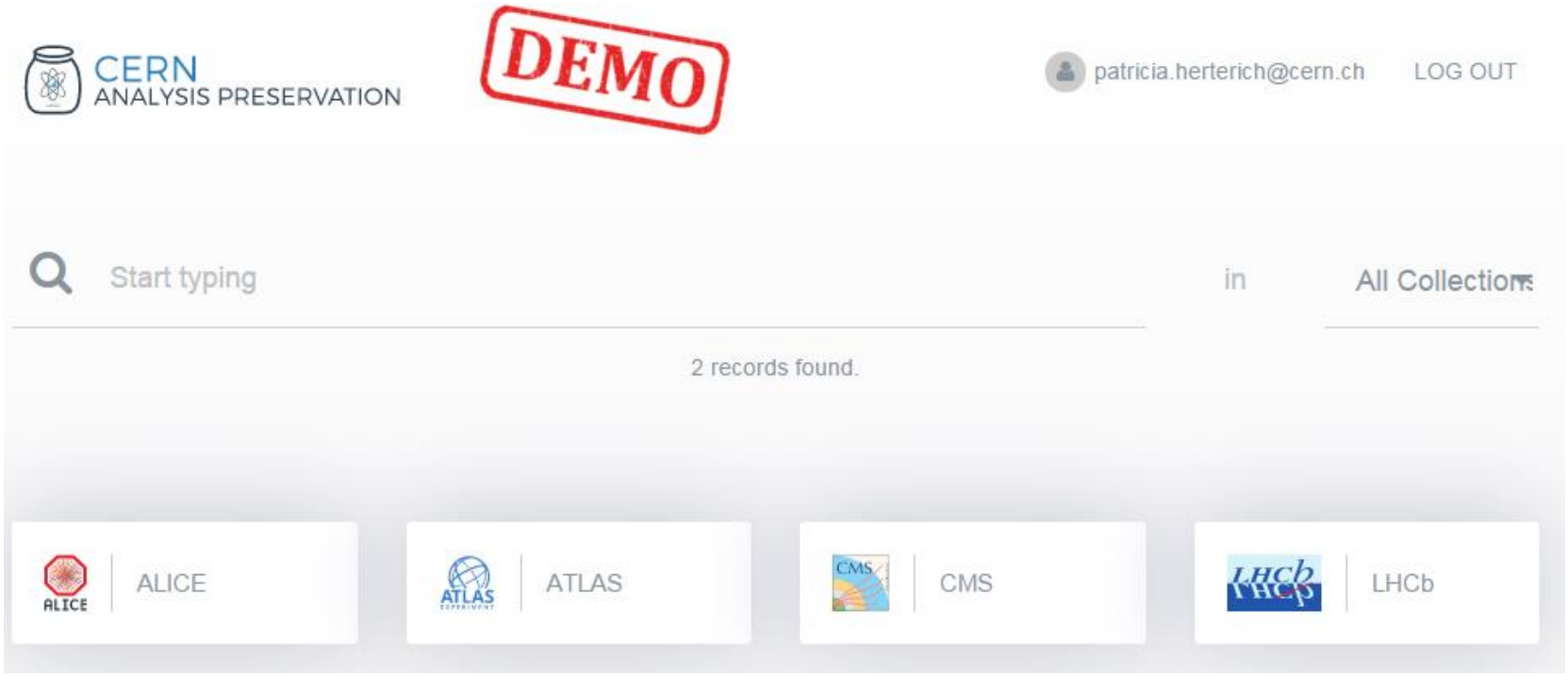
- The data
- The processing steps
- Code involved
- Documentation, Physics information
- Peer review, QA

i.e. all the information contributing to the research claim/presentation/publication to enable future reuse

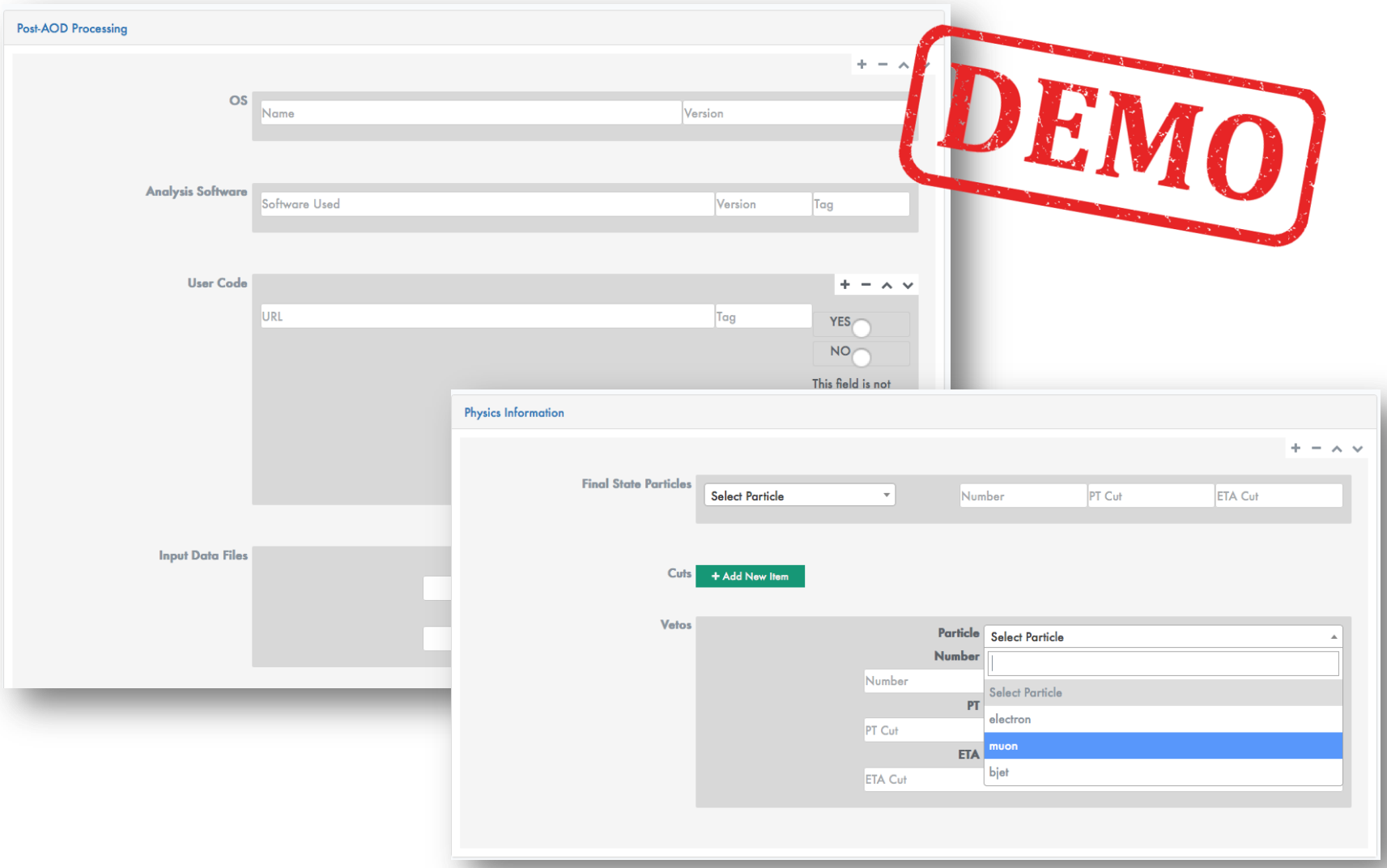
CERN Analysis Preservation



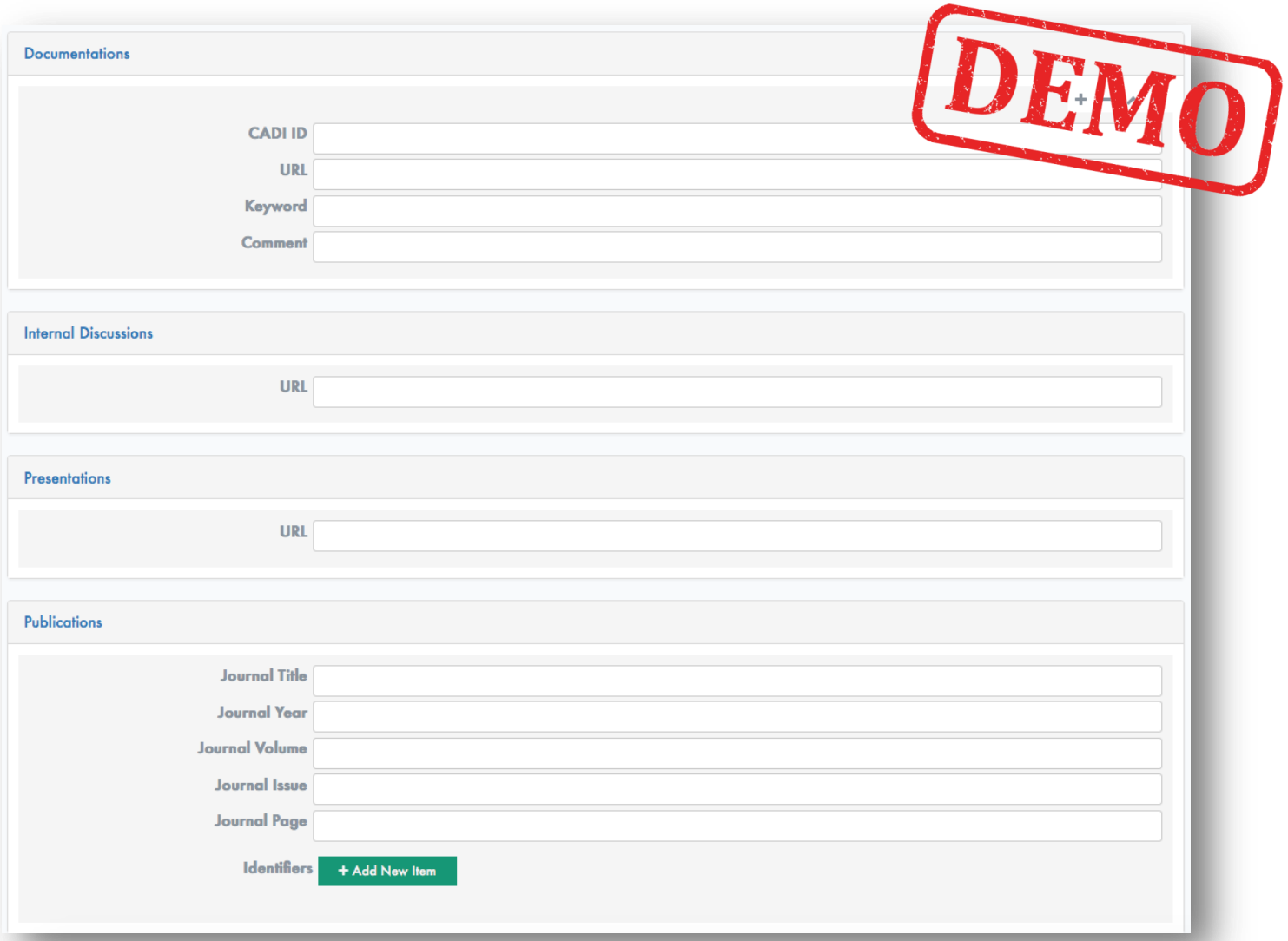
Work in progress



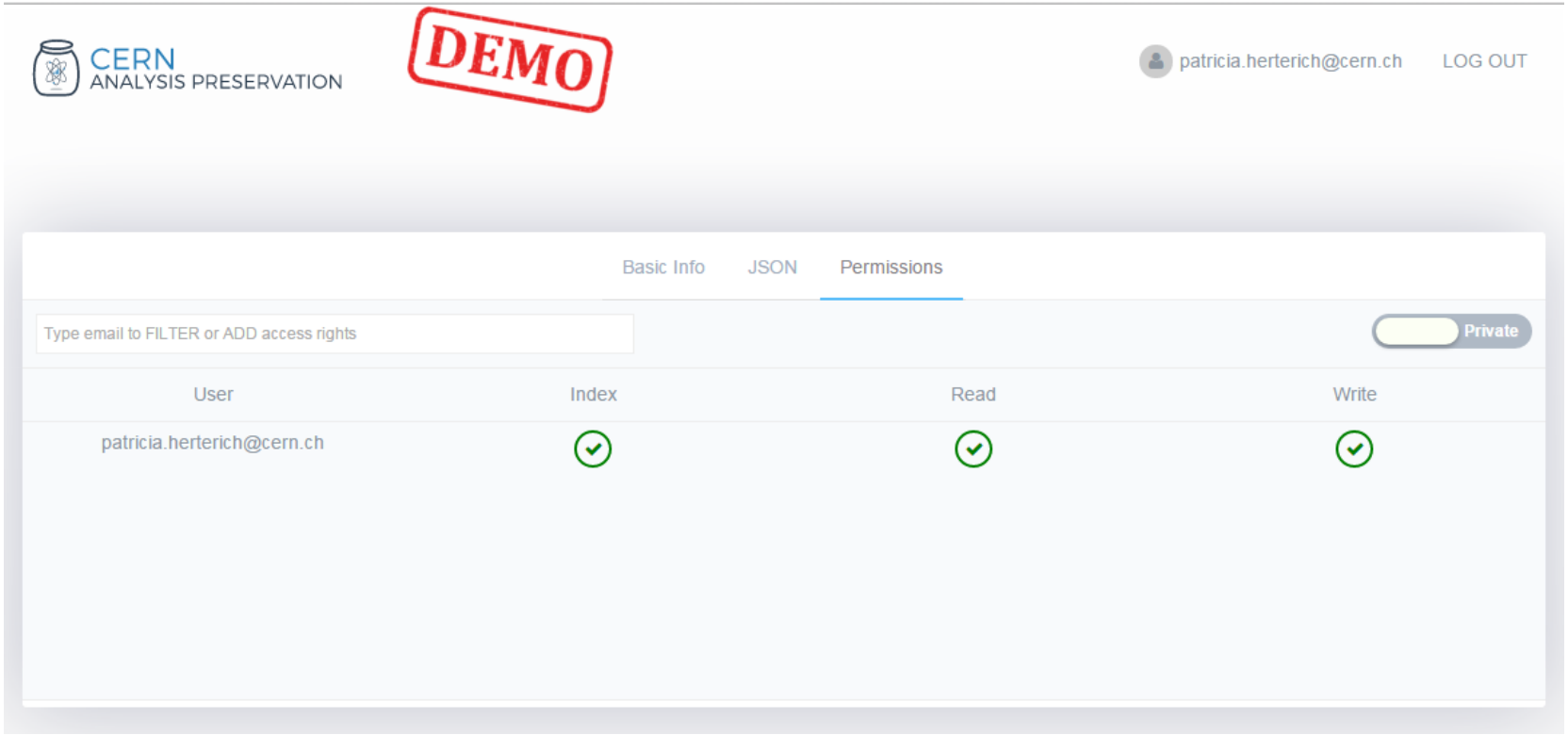
Capturing what is relevant



Integrating other resources



Permissions



Open challenges

- Search syntax for CERN Analysis Preservation
- Integration of CERN Analysis Preservation in publication approval processes, incl. rights and permissions for searching, editing, reviewing
- DOI minting and versioning for content in the CERN Analysis Preservation system
- Publication of preserved analyses, e.g. in the CERN Open Data Portal

Lessons learnt

- Successful services are a collaborative effort between library, IT and the researchers
- Services need to be easy to use for researchers in their everyday workflows and re-use a lot of existing sources
- Don’t be afraid of testing the service often and doing many iterations!

THANKS TO

CERN IT J. Cowton, J. Delgado, J. Kunčar, M. Neumann, T. Smith, T. Šimko

CERN SIS S.Dallmeier-Tiessen, A. Dani, P. Fokianos, L. Rueda

ALICE M. Gheata, C. Grigoras, M. Zimmermann

ATLAS K. Cranmer, L. Heinrich, D. Rousseau, F. Socher

CMS A. Calderon, A. Huffman, K. Lassila-Perini, T. McCauley, A. Rao, A. Rodriguez Marrero

LHCb S. Amerio, M. Bettler, B. Couturier, T. Head, A. Trisovic, A. Ustyuzhanin

CERN CernVM J. Blomer

CERN EOS L. Mascetti

DASPOS M. Hildreth, C. Vardeman, G. Watts

DPHEP F. Berghaus, J. Shiers

Work sponsored by the Wolfgang Gentner Programme of the Federal of Education and Research

SPONSORED BY THE

Federal Ministry of Education and Research

Sources

- CERN convention: http://council.web.cern.ch/council/en/governance/convention.html
- Data policies:
  - ALICE Collaboration (2013). ALICE data preservation strategy. CERN Open Data Portal. http://doi.org/10.7483/OPENDATA.ALICE.54NE.X2EA
  - ATLAS Collaboration (2014). ATLAS Data Access Policy. CERN Open Data Portal. http://doi.org/10.7483/OPENDATA.ATLAS.T9YR.Y7MZ
  - CMS Collaboration (2012). CMS data preservation, re-use and open access policy. CERN Open Data Portal. http://doi.org/10.7483/OPENDATA.CMS.UDBF.JKR9
  - LHCb Collaboration (2013). LHCb External Data Access Policy. CERN Open Data Portal. http://doi.org/10.7483/OPENDATA.LHCb.HKJW.TWSZ