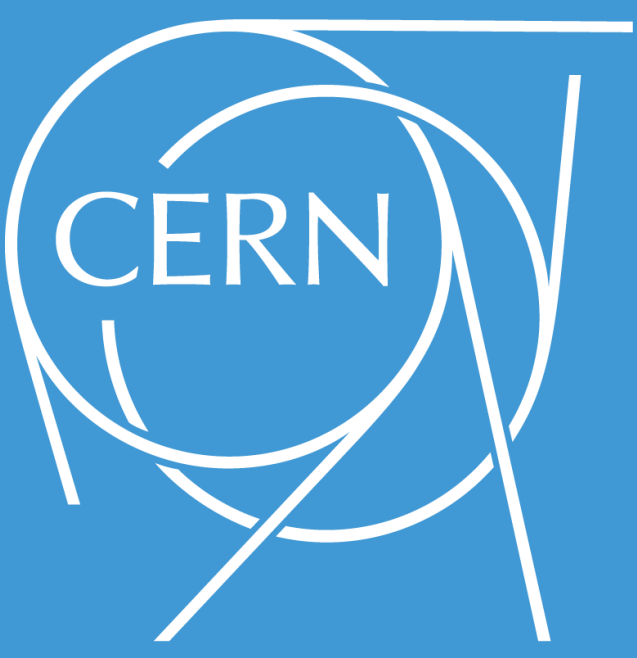


# Python @ CERN

Sebastian Witowski



CERN – home of the Large Hadron Collider that can spit up to **1 petabyte of collision data per second**, impossible for the computing systems to record. The experiments, however, need only a small fraction of “interesting” events. The first-level trigger of ATLAS, for example, selects only **1 in 400 events** (making decisions in ~2.5 millionths of a second). These are sent at **160 GB/second** to the tens of thousands of processor cores which select **1%** of the remaining events for analysis. Even after such drastic data reduction, there are still around **50 petabytes** of data produced at CERN per year.

How does Python fit in this ecosystem? It might not be fast enough to be used for filtering this amount of data, but nevertheless, there are many great projects created with Python at CERN.



Thousands of scientists every day analyse data produced by the experiments at CERN. They need a tool that is based on a flexible and expressive language (like Python), but also works fast with large amount of data. And this is exactly what PyROOT is – a Python module that allows users to interact with ROOT (a data analysis framework written in C++, that is very popular in High Energy Physics community).

PyROOT is more than just a wrapper around ROOT. Python bindings are based on the C++ reflection information stored in the ROOT type system. Thanks to that, Python classes are created dynamically when needed and C++ functions and globals are available automatically in Python.

Users can use PyROOT directly in Python REPL. With all the ROOT classes and functions available out of the box, they can combine them with other modules like NumPy and SciPy to make data analysis easier and faster.

<https://root.cern.ch/pyroot>

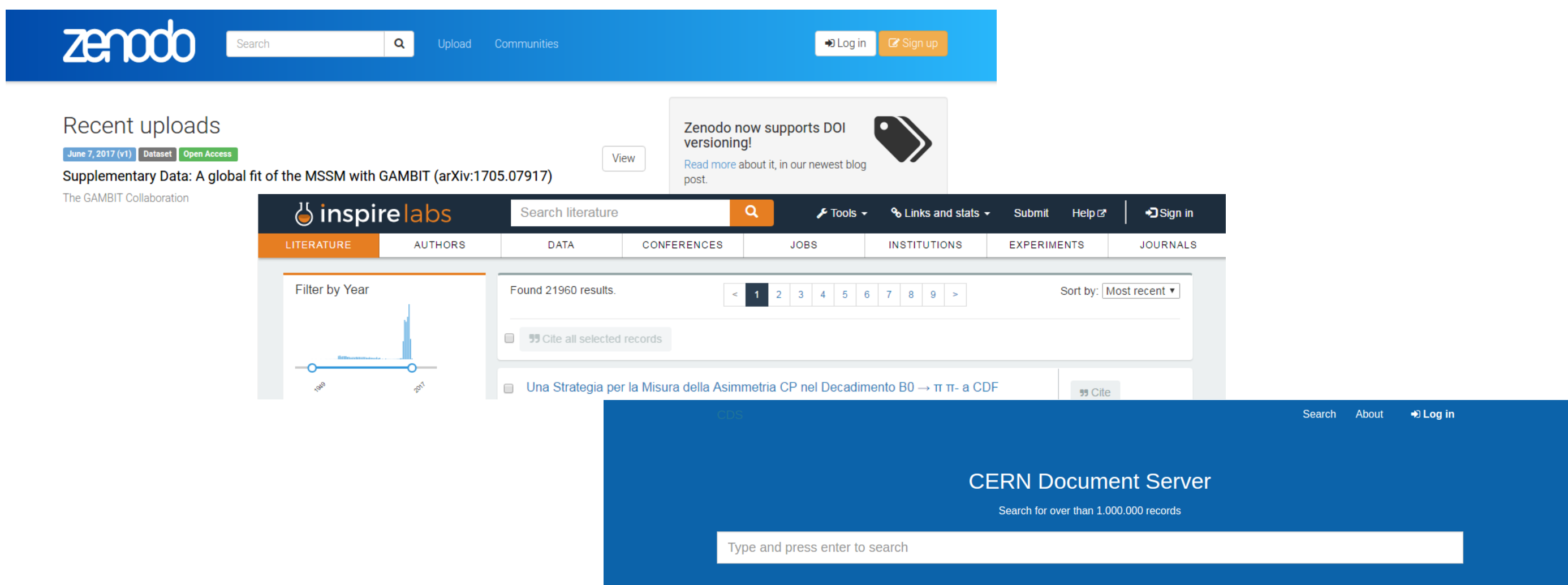


All the discoveries at CERN, the big ones and the small ones, results in thousands of documents that have to go through the publication workflow. For that purpose, a digital library framework called Invenio was created back in 2002.

With this free and open source software you can create:

- Your own integrated digital library system (with module for borrowing books and tools for data curation).
- A multimedia archive or research data repository.
- An institutional repository (with easily defined ingestion and approval workflows to store publications of your institute).

Invenio first started as a monolithic application, but with the recent version 3, it shifted to a more modular approach, built with technologies like Flask, Elasticsearch, Celery or SQLAlchemy.



Invenio powers projects like:

- **CERN Document Server** - CERN’s institutional repository for publications, articles, reports and multimedia content.
- **Inspire** - the main literature database for High Energy Physics, managed by a collaboration of CERN, DESY, Fermilab, IHEP, and SLAC.
- **Zenodo** – a research archive to share and preserve data, software and publications in any size, any format and from any research area (with easy GitHub integration, to make your software citable).



**CERN Analysis Preservation** – a project to capture all the elements needed to understand and rerun an analysis even after several years. Preserving only data is often not enough - once the software becomes obsolete, the research data analyses can no longer be reproduced. CERN Analysis Preservation is trying to solve this problem by collecting not only the data, but also the software, the computing environment, the workflow steps and the documentation.

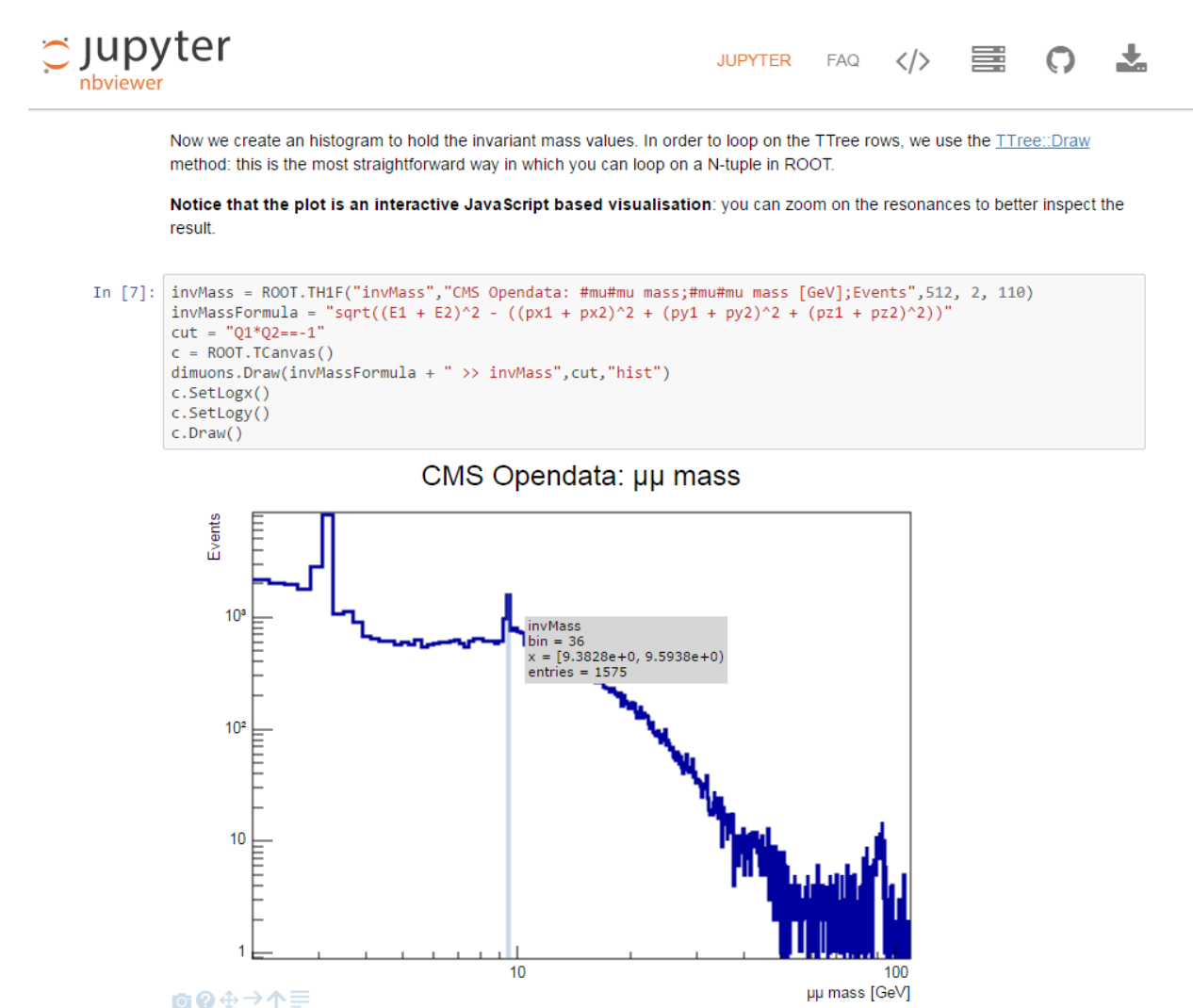


**CERN Open Data** Portal makes the 400TB of data from different experiments at CERN accessible to the world. You can visualize events from the LHC directly in the browser. For more advanced analysis, you can download a Virtual Machine, import the data sets and start your own analysis. And who knows, maybe you will discover something that thousands of scientists missed?

<http://invenio-software.org>



SWAN (Service for Web based Analysis) allows scientists to analyse their data in the CERN cloud. It uses Jupyter notebooks as the interface. The users' sessions are encapsulated in containers, synchronized and stored via CERNBox (file synchronization and sharing service build on top of Owncloud). Large datasets can be accessed on EOS - the disk-based storage system of CERN.



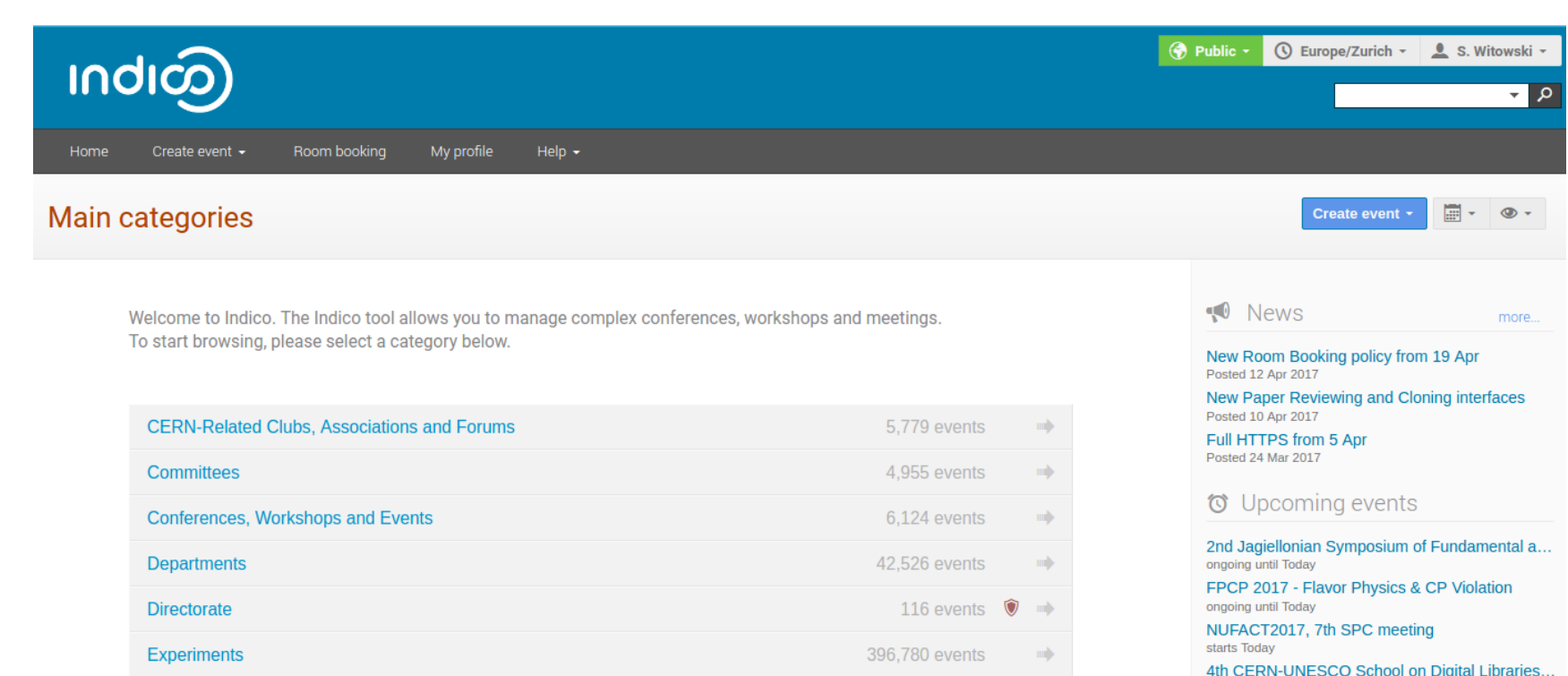
As SWAN is gaining more popularity, a container-based EOS+CERNBox+SWAN solution is being developed. It will allow to deploy SWAN on premises, providing an easy way to setup data analysis environments without the need to install any software.

<https://swan.web.cern.ch>



With thousands of users at CERN (people working on site and visiting scientists), organizing meetings can be a daunting task, but thanks to Indico, it’s actually not that hard.

Indico is an open source tool for event organization, archival and collaboration. You can easily manage the whole event, starting from the registration process, abstract submission, reviewing process up to proceedings submission. It comes with a room booking module, integration with video conferencing tools and few other open source plugins.



Indico at CERN managed its first conference - Computing in High Energy and Nuclear Physics (CHEP) - in 2004 and since then, there have been almost 500 000 events organized with its help. Every week, 25 000 users are using it to organize or participate in meetings.

<http://indico-software.org>

CERN is one of the world’s leading centres for particle physics, but some of the technologies developed here can be applied to other areas, like medical technology, energy, aerospace and more. For example, many diagnostic and therapeutic techniques (like the cancer radiotherapy, which uses beams of protons and other hadron particles) used in the medicine were developed based on the research results obtained at CERN.

Many open source projects are used at CERN and when it's possible, developers try to contribute to those projects. GitLab is used as a version control software, open community versions of MySQL, PostgreSQL and InfluxDB are used by the DB on demand service, Elasticsearch is used throughout many projects. CERN has developed its own Linux distribution called CERN CentOS 7 (based on CentOS and previously on Scientific Linux), its own storage system (EOS), or Dropbox-like cloud synchronization service (CERNBox).

<https://home.cern>

©CERN

CC-BY-SA 4.0

