Python @ CERN

Sebastian Witowski

CERN – home of the Large Hadron Collider that can spit up to **1 petabyte of collisions data per second**. None of today's computing systems are capable of recording such rates, so the first level triggers are selecting only **1 in 10 000 events** which are distributed among **170 computing centers in 42 countries** (the Worldwide LHC Computing Grid) which in turn 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.

python + ROOT Data Analysis Framework

Thousands of scientists every day analyse data produced by the LHC. They need a tool that works fast with gigabytes of data but also is easy to use by someone with little programming experience. 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).

Users can use PyROOT directly in Python REPL. All ROOT classes will be available for them, and on top of that, they can use one of the many great Python modules like NumPy or SciPy to further help them with data analysis.

PyROOT in more than a wrapper around ROOT. Python bindings are based on C++ reflexion information. Thanks to that, Python classes are created dynamically when needed and C++ functions and globals are available automatically in Python.

https://root.cern.ch/pyroot

SWAN

SWAN or Service for Web based Analysis is a platform to perform interactive data analysis in the cloud. Currently tested at CERN, SWAN creates Virtual Machines that automatically connect CERNBox (cloud storage built on top of Owncloud that uses EOS as a backend) with Jupyter notebook interface.

SWAN allows users to:
- Access the data and software available on the connected file system
- Easily store their own results and share them with others
- Run shell commands in the VM from the browser

Currently in the alpha phase, the container-based EOS+CERNBox+SWAN services will provide an easy way to analyse data without the need to install any software.

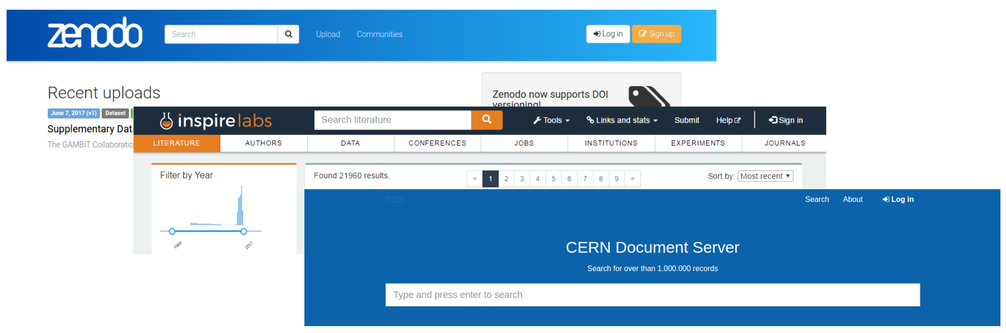
https://swan.web.cern.ch/

INVENIO

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 Invenio was created back in 2002.

With this free an open source software you can create:
- Your own integrated library system (with module for borrowing books and tools for data curation)
- A multimedia archive
- 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.



Invenio powers projects like:
- CERN Document Server - CERN's official repository for publications, articles, reports and multimedia content.
- Inspire - the biggest database for High Energy Physics literature run by a collaboration of CERN, DESY, Fermilab, IHEP, and SLAC.
- Zenodo – an open access repository to share and preserve data, software and publications in any size from any research area (with easy GitHub integration, you can make your software citable).

CERN Analysis Preservation

- CERN Analysis Preservation – a project to preserve data from various experiments together with the software and worflows to make them reproducible even after many years.

opendata CERN

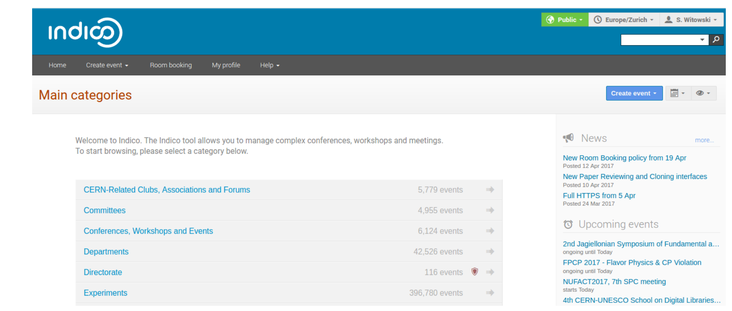
- CERN Open Data Portal stores data from different experiments at CERN. You can visualize events directly in the browser and for more advanced analysis, you can create a Virtual Machine, import the data set and start your own analysis.

http://invenio-software.org

indico

With thousands of users at CERN (people working on the 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 it's first conference - CHEP - in 2004 and since then, there have been almost 500 000 events organized with it's help. Every week, 25 000 users are using it to organize or participate in meetings.

http://indico-software.org

CERN – home of the Large Hadron Collider that can spit up to **1 petabyte of collisions data per second**. None of today's computing systems are capable of recording such rates, so the first level triggers are selecting only **1 in 10 000 events** which are distributed among **170 computing centers in 42 countries** (the Worldwide LHC Computing Grid) which in turn select 1% of the remaining events for analysis. Even after such drastic data reduction, there are still