# Digital Repositories at CERN

Jose Benito Gonzalez Lopez



## Who are we?

SW Eng. team providing DR technology



Develop software and provide services

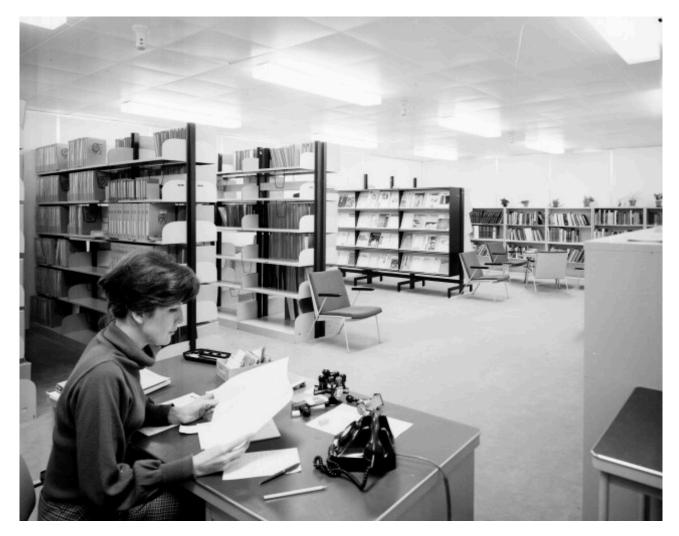


Open: Source, Data, Science





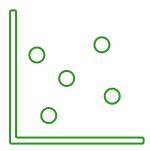
# Repository?



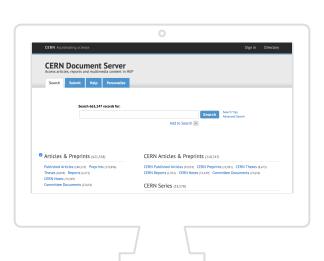


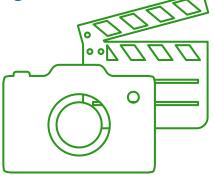


Articles,
Papers,
Thesis,
Standars,
Books, etc



Data, Software, ...





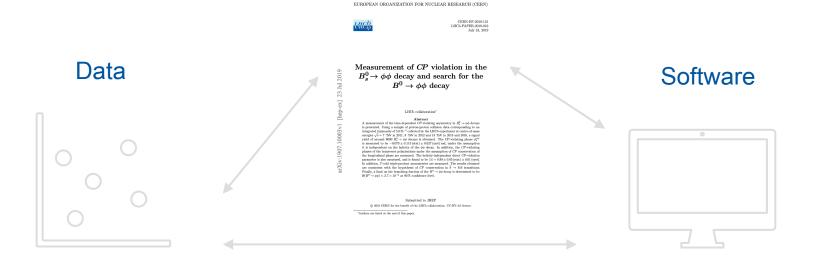
Photos, Videos, Audio, ...





# Open Science

## **Publication**



De-duplication
Sharing
"Fake science"
Effective research





















## Publish or Perish?

20%

store data in a digital archive





ZENODO

DAIS ARCHIVAL STORE

CERN ANALYSIS PRESERVATION

INSPIRE

HEP DATA, SCOPE3

60 INSTANCES WORLD WIDE

INVENIO)





## Invenio

https://inveniosoftware.org http://github.com/inveniosoftware





## Invenio













Born at **CERN** Free Open Source Software **Core** for any digital repository **Modern** & reliable technology Flexible and modular Handling 100M+ records Develop with **PB** files in mind Fast upload/search



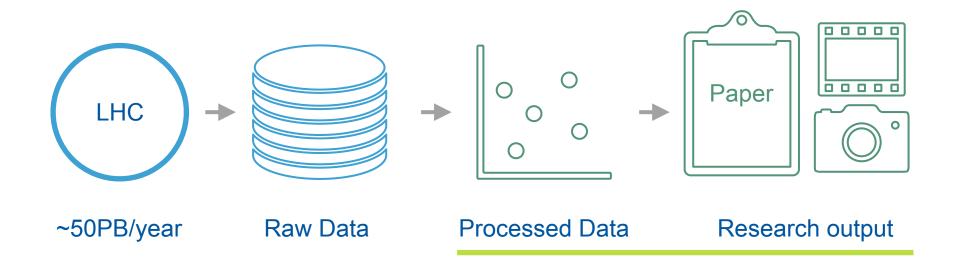
INVENIO



## Invenio Services at CERN



## Data at CERN

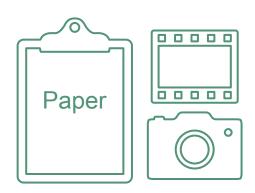




## At CERN







Research output

CERN
ANALYSIS
PRESERVATION

**REANA** 

CERN OPEN DATA

SERVER

**CERN DOCUMENT** 

http://analysispreservation.cern.ch

https://reana.io

http://opendata.cern.ch

http://cds.cern.ch

Capture

Reproduce

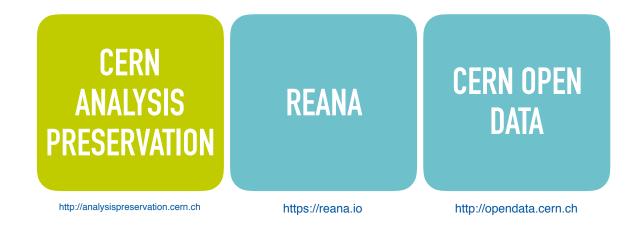
**Publish** 

**Publish and Archive** 



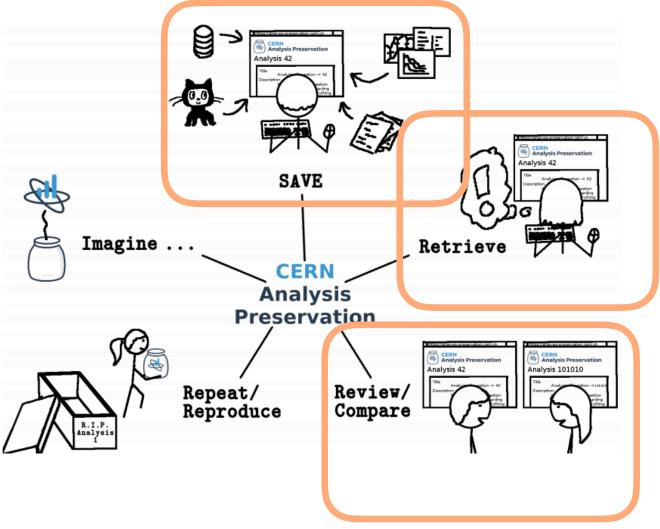


## **Processed Data**





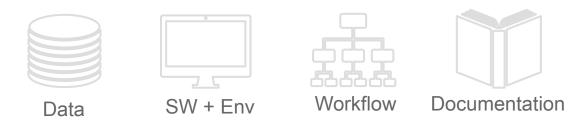
# **CERN Analysis Preservation**

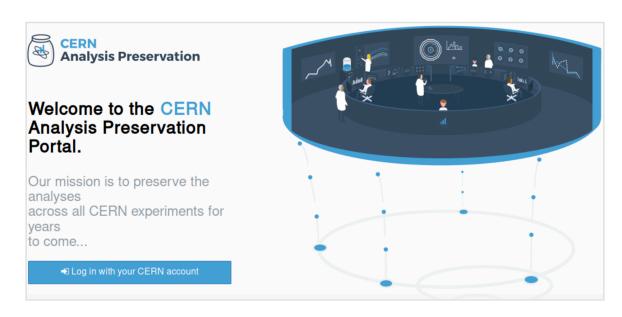




# **CERN Analysis Preservation**

Capturing all the elements needed to understand and rerun an analysis even several years





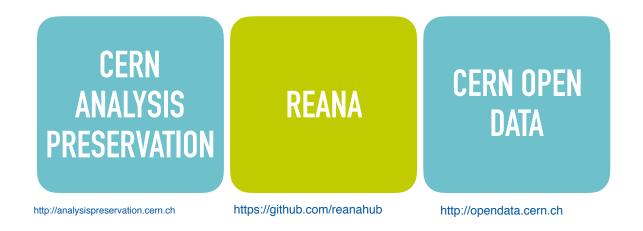
http://analysispreservation.cern.ch

https://github.com/cernanalysispreservation





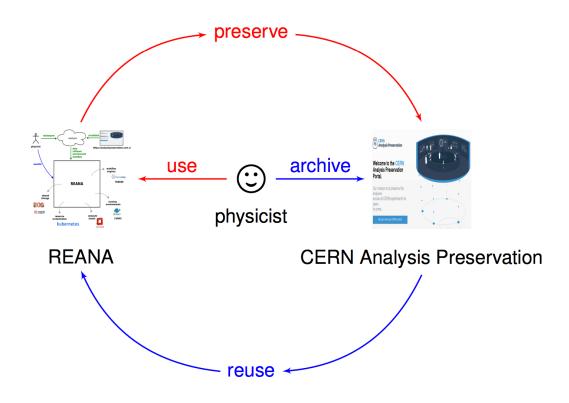
## **Processed Data**





# RE(usable)ANA(lysis)

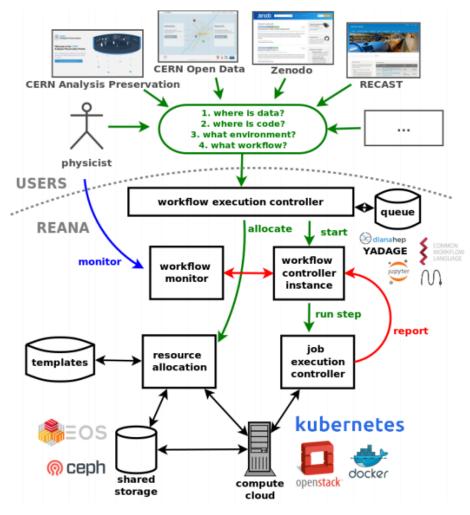
https://github.com/reanahub





# RE(usable)ANA(lysis)

https://github.com/reanahub







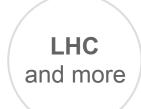
## **Processed Data**





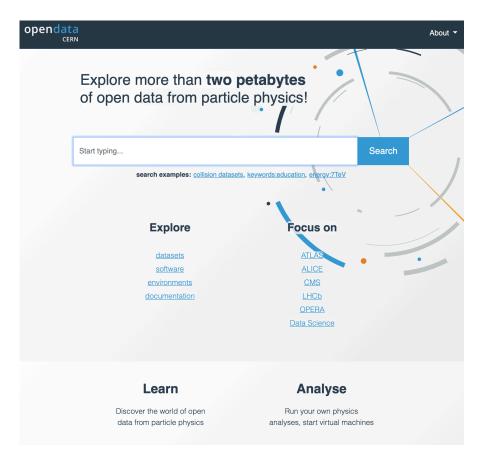
# **CERN Open Data Portal**

Publicly-accessible site for curated releases of CERN data sets and software



2016 CMS 300 TB

2017 CMS ~1 PB



2019 CMS Data S.

http://opendata.cern.ch

https://github.com/cernopendata











## BTag primary dataset in AOD format from RunA of 2011 (/BTag/Run2011A-12Oct2013-v1/AOD) 2016

/BTag/Run2011A-12Oct2013-v1/AOD CMS collaboration

Cite as: CMS collaboration (2016). BTag primary dataset in AOD format from RunA of 2011 (/BTag/Run2011A-12Oct2013-v1/AOD). CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.N372.QF6S

Collection: CMS-Primary-Datasets Collision Energy: 7TeV Experiment: CMS Accelerator: CERN-LHC Parent Dataset:

/BTag/Run2011A-v1/RAW

## Description

BTag primary dataset in AOD format from RunA of 2011. Run period from run number 160404 to 173692.

### **Notes**

This dataset contains all runs from 2011 RunA. The list of validated runs, which must be applied to all analyses, can be found in

### None

## Characteristic 3

Dataset: 11759539 ever s 489 files 1.8 TB in total

## System Details

Global tag: FT\_53\_LV5\_AN1

Recommended release for analysis: CMSSW\_5\_3\_32

### How were these data selected?

Dataset defined for the calibration of b-quark tag algorithms. Events stored in this primary dataset were selected because of the presence of at least two high-energy jets, where one of them is tagged as a b-quark jet with a soft muon from the b-quark decay in the event.

### Data taking / HLT

The collision data were assigned to different RAW datasets using the following HLT configuration.

## Export

**JSON** 

The collision data were assigned to different RAW datasets using the following HLT configuration.

### Data processing / RECO

This primary AOD dataset was processed from the RAW dataset by the following step:

Step: RECO

Release: CMSSW\_5\_3\_12\_patch1 Global tag: FT\_R\_53\_LV5::All

Configuration file for RECO step reco\_2011A\_BTag

### **HLT** trigger paths

The possible HLT trigger paths in this dataset are:

HLT\_BTagMu\_DiJet110\_Mu5 HLT\_BTagMu\_DiJet20\_Mu5 HLT\_BTagMu\_DiJet40\_Mu5

HLT\_BTagMu\_DiJet70\_Mu5

### How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

CMS data quality monitoring: Systems and experiences

The CMS Data Quality Monitoring software experience and future improvements

The CMS data quality monitoring software: experience and future prospects

## How can you use these data?

You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in

How to install the CMS Virtual Machine

Getting started with CMS open data

### File Indexes

Filename	Size	Download	EOS Link

# Filename Size Download EOS Link CMS\_Run2011A\_BTag\_AOD\_12Oct2013-v1\_00000\_file\_index.txt 122 Bytes ★ ② CMS\_Run2011A\_BTag\_AOD\_12Oct2013-v1\_20000\_file\_index.txt 59.5 kB ★ ②

### FirstPrevious1NextLast

File Indexes

### **Datasets**

Filename	Size	Download	EOS Link
802CF580-BB46-E311-8D89-00261894388D.root	886.7 MB	<u>+</u>	0
00376186-543E-E311-8D30-002618943857.root	3.9 GB	<u>+</u>	0
0080432E-043E-E311-B4CB-00248C0BE01E.root	2.8 GB	<u>+</u>	0
00867474-453E-E311-A450-003048FFD7C2.root	3.9 GB	<u>+</u>	e
02012C2B-323E-E311-897E-003048FFD736.root	2.2 GB	<u>+</u>	e
02116E88-003E-E311-A1A9-0025905964BA.root	4.2 GB	<u>+</u>	0
0216066B-3A3E-E311-ABD0-003048FFD732.root	3.9 GB	<u>+</u>	0
02477509-3D3E-E311-A230-00261894389A.root	3.9 GB	<u>+</u>	0
02581093-3E3E-E311-8235-00248C55CC3C.root	3.9 GB	<u>+</u>	0
0297C037-2D3E-E311-83A2-00259059649C.root	4.2 GB	<u>+</u>	0

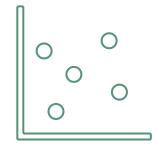
### FirstPrevious12345NextLast

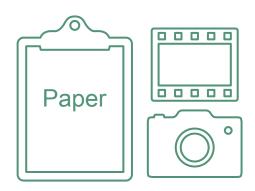
### Disclaimer

The open data are released under the Creative Commons CC0 waiver. Neither CMS nor CERN endorse any works, scientific or otherwise, produced using these data. All releases will have a unique DOI that you are requested to cite in any

## At CERN







**Processed Data** 

Research output



http://analysispreservation.cern.ch

**REANA** 

https://reana.io

CERN OPEN DATA

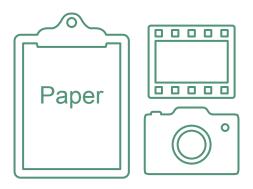
http://opendata.cern.ch

CERN DOCUMENT SERVER

http://cds.cern.ch





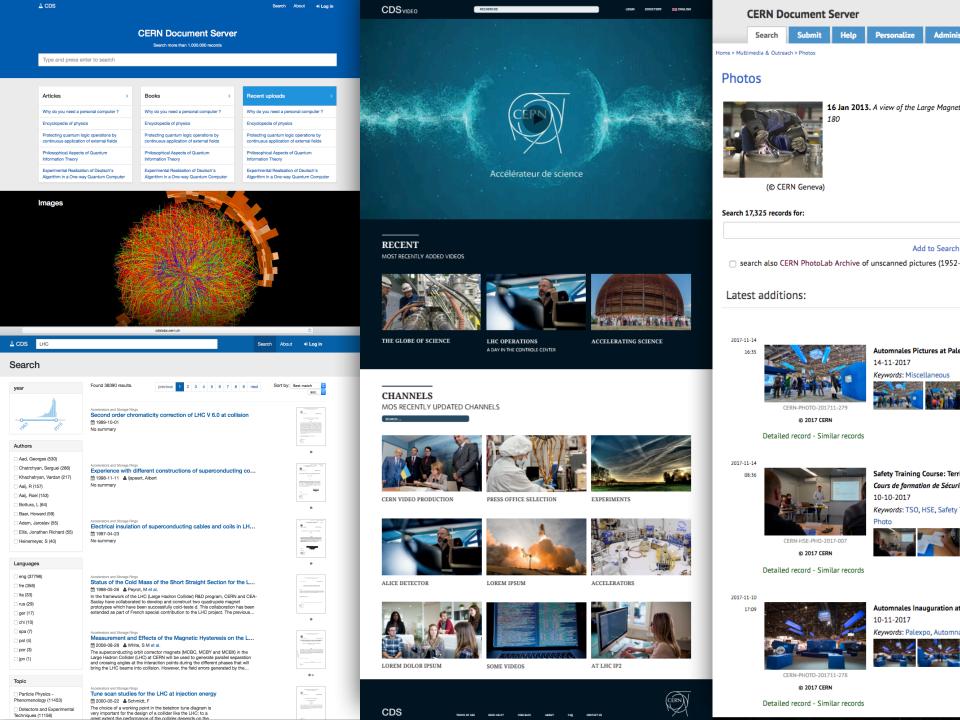


## Research output

# CERN DOCUMENT SERVER

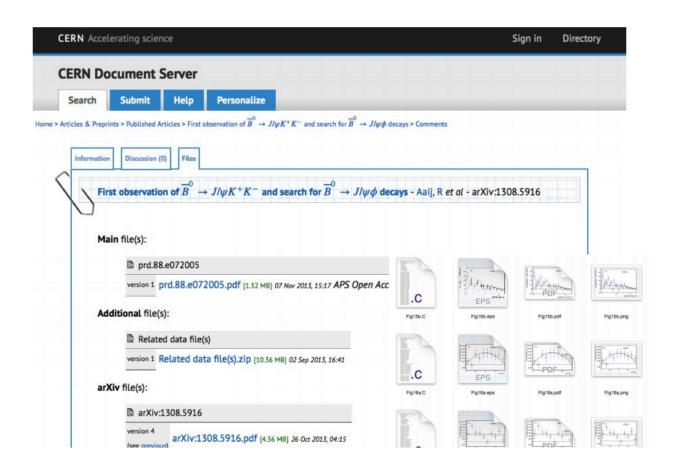
http://cds.cern.ch





## CDS

## Also used for small data







# **Beyond CERN**





# **Beyond CERN**

### Zenodo





Long tail of science

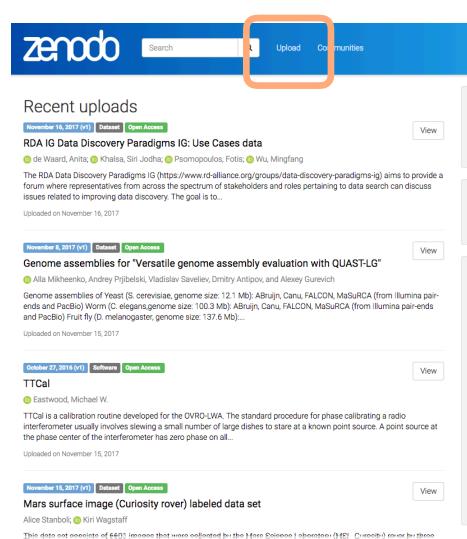
(+) 50GB uploads

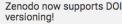
Running on latest Invenio tech

Using CERN Computing Centre

Using EOS Storage Technology





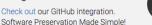




♣ jose.benito.gonzalez@cern.ch

Read more about it, in our newest blog post.

### Usina GitHub?





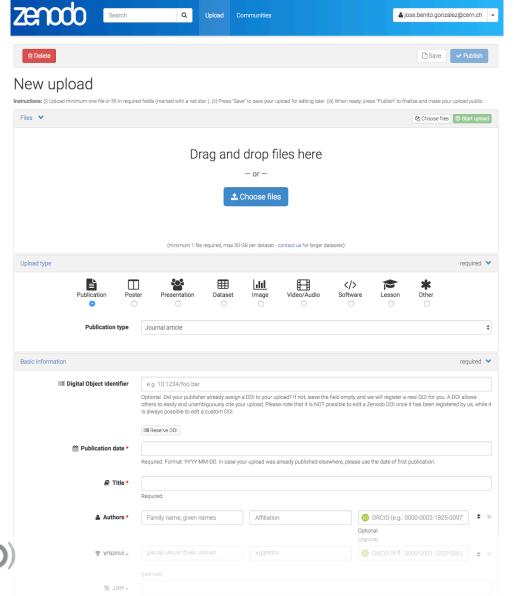
### Zenodo in a nutshell

- · Research. Shared. all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
- Citeable. Discoverable. uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- · Communities create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
- · Funding identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- Flexible licensing because not everything is under Creative Commons.
- Safe your research output is stored safely for the future in the same cloud infrastructure as CERN's own LHC research data.

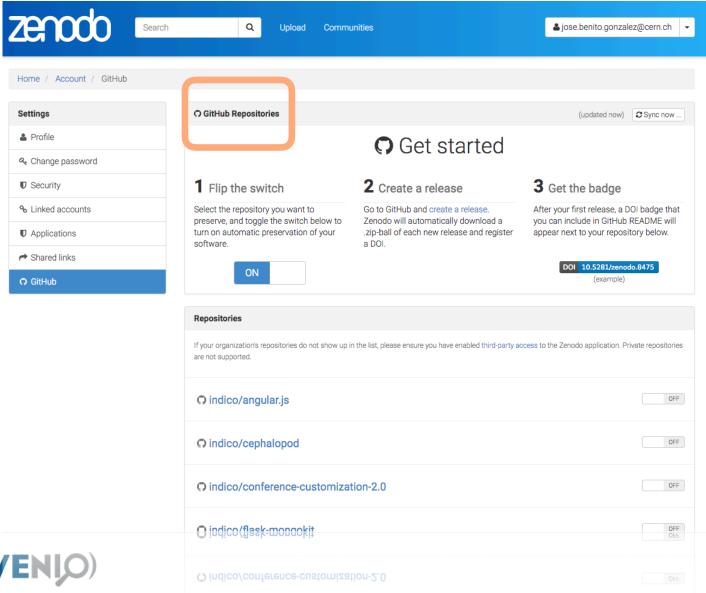
Read more about Zenodo and its features.

















November 18, 2017

Software Open Access

## ligo-cbc/pycbc: post-02 release 3

Alex Nitz; Ian Harry; Duncan Brown; Christopher M. Biwer; Josh Willis; Tito Dal Canton; Larne Pekowsky; Thomas Dent; Andrew R. Williamson; Collin Capano; Soumi De; Miriam Cabero; Bernd Machenschalk; Prayush Kumar; Steven Reyes; Thomas Massinger, Amber Lenon; Stephen Fairhurst; Alex Nielsen; shasvath; Francesco Pannarale; Leo Singer; Duncan Macleod; Stanislav Babak; Hunter Gabbard; John Veitch; CBC Sugar; Sebastian Khan; dfinstad; Lorena Magaña Zertuche

This is the third post-O2 release of PyCBC for analysis of data taken during Advanced LIGO's second observing run and Advanced Virgo's first observing run.

This release has been tested against LALSuite with the hash:

95ad957cee1a37b7fc3128883d8b723556f9ec38

This release updates the qtransform feature set and adds an interface to access data and information about the catalog of gravitational wave mergers.

Details of the changes since the last release are at https://github.com/ligo-cbc/pycbc/compare/v1.8.1...v1.8.2

A Docker container for this release is available from the pycbc/pycbc-el7 repository on Docker Hub be downloaded using the command:

### docker pull pycbc/pycbc-el7:v1.8.2

bank

On a machine with CVMFS installed, a pre-built virtual environment is available for Red Hat 7 compatible operating systems by running the command:

source /cvmfs/oasis.opensciencegrid.org/ligo/sw/pycbc/x86\_64\_rhel\_7/virtualenv/pycbcv1.8.2/bin/activate

and for Debian 8 compatible operating systems by running the command:

source /cvmfs/oasis.opensciencegrid.org/ligo/sw/pycbc/x86\_64\_deb\_8/virtualenv/pycbcv1.8.2/bin/activate

A bundled pycbc\_inspiral executable for use on the Open Science Grid is available at

/cvmfs/oasis.opensciencegrid.org/ligo/sw/pycbc/x86\_64\_rhel\_6/bundle/v1.8.2/pycbc\_inspiral





### **Publication date:**

November 18, 2017

DOI:

DOI 10.5281/zenodo.1058970

### Related identifiers:

Supplement to:

https://github.com/ligo-cbc/pycbc/tree/v1.8.2

### License (for files):

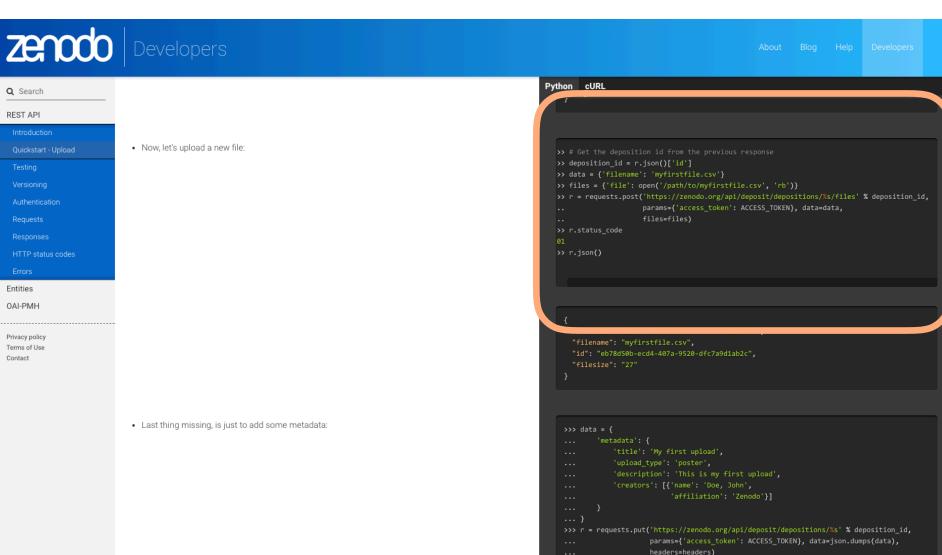
Other (Open)

Version v1.8.2 10.5281/zenodo.1058970	Nov 18, 2017			
Version v1.8.1 10.5281/zenodo.888262	Sep 10, 2017			
Version v1.8.0 10.5281/zenodo.887622	Sep 8, 2017			
Version v1.7.11 10.5281/zenodo.883086	Sep 1, 2017			
Version v1.7.10 10.5281/zenodo.852372	Aug 28, 2017			
View all 48 versions				





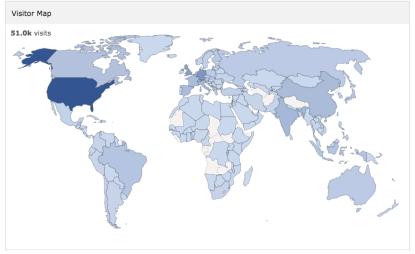
http://developers.zenodo.org/



50K Monthly visitors

30 TB Stored Files









# Are DR important?

- Open Science is about de-duplication and efficiency
- We need to store all research objects properly
  - SW, data, documentation, articles, etc
- Journals publish those articles and start linking to related objects
  - 5 publishers control ~70% of market
  - Profits are bigger than Google and co
- Good Research depends on Open Accessible Science
- Invenio helps building repositories of all kinds



# Challenges

- Replicate published work = "Crisis"
  - Wider sharing and reanalysis of code, data, and research materials...All objects have persistent identifiers
    - interactive and more transparent ways of presenting data graphically (e.g. building graphs)
- Open access publishing (breaking The Wall)
- Interoperability between repositories (COAR and NGR)
- Predatory publishers
  - How to differentiate good research from fake research
- Mapping research? Where? Who? how?



## How we can help making open science better and easier?

Please, solve this:)



## Questions?



https://inveniosoftware.org

http://github.com/inveniosoftware

https://invenio.readthedocs.io/en/latest/

7FMNNN 7

32SHARE

**OAIS ARCHIVAL STORE** 

REANA

SPIRE, HEP DATA, SCOPE

O INSTALLATIONS
WORLD WIDE

INVENIO)





## **OAI Archival Store**

Long term preservation
Using Archivematica (FOSS)
Integration done through Invenio module
Intend to integrate with CDS & Zenodo &...

https://github.com/CERN-E-Ternity

