TOOLS & SERVICES FOR OPEN SCIENCE

Ludmila Marian
Scientific Data Manager, IAEA
WHAT IS NOT OPEN SCIENCE?

Published Paper

Software
Data
Other research objects
OPEN THE RESEARCH PROCESS

- Clear citations
- Ensure the accumulation of credits
- Publish metadata with an open licence
- Use open evaluation
- Ensure links between publications, data and methods
- Make use of institutional repositories
- Use services that safeguard the preservation and integrity of materials
- Produce standard metadata
- Consider financiers’ requirements
- Clarify usage rights
- Ensure that you give credit through citations
- Make use of open-source software and open interfaces
- Make use of service infrastructure
- Attach a persistent identifier to your results
- Attach descriptive metadata to your results
- Publish metadata with an open licence

https://www.fosteropen-science.eu/content/what-open-science-introduction
OPEN RESEARCH
VALIDATE AND REUSE RESEARCH

- **Paper**
- **Data**
- **Software**
- **Other research objects**

Rerun
WHERE TO STORE THESE OBJECTS?
WHERE TO STORE THESE OBJECTS?

In a Digital Repository!
WHY A DIGITAL REPOSITORY?
PUBLISH OR PERISH?

20%

store data in a digital archive
Open Source framework for large-scale digital repositories.
ABOUT INVENIO

• Born at **CERN**

• **Free Open** Source Software

• **Core** for any digital repository

• **Modern** & reliable technology

• **Flexible** and modular

• Handling **100M+** records

• Developed with **PB** in mind

• **Fast** uploads & search
DATA @ CERN

1. Provision of additional documentation for the published results
2. Simplified data formats for analysis in outreach and training exercises
3. Reconstructed data and simulations as well as the analysis level software to allow a full scientific analysis
4. Basic raw level data (if not yet covered as level 3 data) and their associated software which allows access to the full potential of the experimental data

LHC
~50PB/year
Raw Data
Processed Data
Research output
Paper
DATA @ CERN

Processed Data

CERN ANALYSIS PRESERVATION
http://analysispreservation.cern.ch

REANA
https://reana.io

CERN OPEN DATA
http://opendata.cern.ch

CERN DOCUMENT SERVER
http://cds.cern.ch

Capture
Reproduce
Publish
Publish and Archive

Research output

Paper
Processed Data

CERN ANALYSIS PRESERVATION
http://analysispreservation.cern.ch

REANA
https://reana.io

CERN OPEN DATA
http://opendata.cern.ch
CERN ANALYSIS PRESERVATION

Imagine ...

SAVE

Retrieve

Repeat/Reproduce

Review/Compare
Capturing all the elements needed to understand and rerun an analysis even several years later

Data  SW + Env  Workflow  Documentation
CODE + DATA IS NOT ENOUGH

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild, Petra Habets, Heidi L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Published: June 1, 2012 • DOI: 10.1371/journal.pone.0038234

Software changes: 8.8±6.6% (volume) and 2.8±1.3% (thickness)
Operating system changes: factor two smaller
THE FULL RESEARCH MUST BE CAPTURED
REPRODUCIBILITY IS HARD

Half of researchers cannot reproduce their own results

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
Processed Data

CERN ANALYSIS PRESERVATION
http://analysispreservation.cern.ch

REANA
https://reana.io

CERN OPEN DATA
http://opendata.cern.ch
Reproducible analysis

Reproducible research data analysis platform

Flexible
Run many computational workflow engines.

Scalable
Support for remote compute clouds.

Reusable
Containerise once, reuse elsewhere. Cloud-native.

Free
Free Software. MIT licence. Made with ❤ at CERN.

reana.io
REPRODUCIBLE WORKFLOWS

Serial

Yadage

CWL

inputs

command₁

output₁

command₂

output₂

...

outputₙ₋₁

commandₙ

outputs
REANA
Processed Data

CERN ANALYSIS PRESERVATION
http://analysispreservation.cern.ch

REANA
https://reana.io

CERN OPEN DATA
http://opendata.cern.ch
CERN OPEN DATA PORTAL

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

Explore more than **two petabytes**
of open data from particle physics!

search examples: collision datasets, keywords:education, energy:7TeV

Explore
- datasets
- software
- environments
- documentation

Focus on
- ATLAS
- ALICE
- CMS
- LHCb
- OPERA
- Data Science

2 PB data
CERN OPEN DATA PORTAL

Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD)

CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.BMAR.CAA2

Description
Mu primary dataset in AOD format from RunB of 2010

Notes
This dataset contains all runs from 2010 RunB. The list of validated runs, which must be applied to all analyses, can be found in
CMS list of validated runs Cart_136003-149442_77eV_Apr21ReReco_Collisions10_f2011_v2.txt

Related Datasets
/Mu/Run2010B-v1/RAW

Characteristics
Dataset: 3237691 events 2979 files 3.2 TB in total

System Details
Global tag: FT_R.42_V10A:Ali
Recommended release for analysis: CMSSW_4.2.1_patch1

How were these data selected?
There are four categories of triggers in the Mu dataset (with significant overlaps):
• 70% inclusive single muon triggers with varying trigger pt threshold 3.5, 5, 7, 9, 11, 13, 15, 17, 19, 21 GeV plus a few with loosened quality cuts.
• 20% isolated single muon triggers with varying trigger pt threshold 9, 11, 13, 15, 17 GeV.
• 10% inclusive dimuon triggers with varying trigger pt threshold 3.5 GeV plus one Z->mu+mu trigger with loosened quality cuts.
• 20% combinations of muon triggers with various pt thresholds 3.5, 7, 8, 9, 11 GeV with some EM/gamma or hadronic/jet energy deposit with thresholds 6-100 GeV.

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: 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
### Datasets

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Download</th>
<th>EOS Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS_Run2011A_BTag_AOD_12Oct2013-v1_00000_file_index.txt</td>
<td>122 Bytes</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>CMS_Run2011A_BTag_AOD_12Oct2013-v1_20000_file_index.txt</td>
<td>59.5 kB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>802CF580-BB46-E311-8D89-00261894388D.root</td>
<td>886.7 MB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>00376186-543E-E311-8D30-002618943857.root</td>
<td>3.9 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>0080432E-043E-E311-B4CB-00248C0BE01E.root</td>
<td>2.8 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>00867474-453E-E311-A450-003048FFD7C2.root</td>
<td>3.9 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>02012C2B-323E-E311-897E-003048FFD736.root</td>
<td>2.2 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>02116E88-003E-E311-A1A9-0025905964BA.root</td>
<td>4.2 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>0216066B-3A3E-E311-ABD0-003048FFD732.root</td>
<td>3.9 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>02477509-3D3E-E311-A230-00261894389A.root</td>
<td>3.9 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>02581093-3E3E-E311-8235-00248C55CC3C.root</td>
<td>3.9 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
<tr>
<td>0297C037-2D3E-E311-83A2-00259059649C.root</td>
<td>4.2 GB</td>
<td><img src="download.png" alt="Download" /></td>
<td><img src="eos.png" alt="EOS Link" /></td>
</tr>
</tbody>
</table>

### 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 applications or publications.
CERN OPEN DATA PORTAL
Higgs-to-four-lepton analysis example using 2011-2012 data

Jonhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Azuddin;
Cite as: Jonhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Azuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JR8B.RR42

Description

This research level example is a strongly simplified reimplemention of parts of the original CMS Higgs to four lepton analysis published in Phys.Lett. B716 (2012) 30-61, arXiv:1207.7235.

The published reference plot which is being approximated in this example is https://inspirehep.net/record/1124338/files/H4L_mass_3.png. Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered by this example.

The example consists of different levels of complexity. The highest level is a minimal understanding of the content of this paper and of the main educational exercises. The lower levels might also be interesting for educational purposes.

Use with

The example uses legacy versions of the original CMS datasets in the Atoms publication due to improved calibrations. It also uses legacy versions of the analysis but not identical to the ones in the original publication. These legacy versions are in many later CMS publications.

/DoubleMu/Run2011A-12Oct2013-v1/AOD
Independent analyses by theorists (Jesse Thaler et al, MIT)
LONG TERM VALUE OF DATA

Collaborations publish papers even ~15 years after data taking ends

JADE data (1979–1986) still unique even ~35 years later
Research output

CERN DOCUMENT SERVER

http://cds.cern.ch

Publish and Archive
CERN DOCUMENT SERVER
ALSO USED FOR SMALL DATA

- 2.3 Million records
- 6.5 Million files
- 140 TB of data
- 5’000 unique visits / day
Zenodo
http://zenodo.org

Catch-all for the long-tail of science

<table>
<thead>
<tr>
<th>Recent uploads</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Colour 0.3.14</strong></td>
</tr>
<tr>
<td>Mansencal, Thomas; Mauderer, Michael; Parsons, Michael; Shaw, Nick; Wheatley, Kevin; Cooper, Sean; Vanderberg, Jean-D.; Caneran, Luke; Crowson, Katherine; Lex, Olve; Leenweber, Katrin; Sharma, Shriraman; Sobotka, Troy; James; Moritz, Dominik; Pipp, Matt; Rane, Chirn; Esawi-Moorthi, Paithri; Merrit, Jon; Paleolithic, Ber; Leonard, Manuel</td>
</tr>
<tr>
<td>Colour Science for Python Colou, is an open-source Python package providing a comprehensive number of algorithms and datasets for colour science. It is freely available under the New BSD License terms. Colour is an affiliated project of NumFOCUS, a 501(c)(3) nonprofit in the United States...</td>
</tr>
<tr>
<td>Uploaded on October 29, 2019</td>
</tr>
<tr>
<td>I'd more version(s) exist for this record</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recent uploads</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistics and Evaluation Data for Publication &quot;Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research&quot;</strong></td>
</tr>
<tr>
<td>Tobias Weber; Michael Fromm; Nelson Tavares de Sousa</td>
</tr>
<tr>
<td>Automated classification of metadata of research data by their discipline(s) of research can be used in scientometric research, by repository service providers, and in the context of research data aggregation services. Openly available metadata of the DataCite Index for research data were used...</td>
</tr>
<tr>
<td>Uploaded on October 15, 2019</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recent uploads</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene Ontology Data Archive</strong></td>
</tr>
<tr>
<td>Carbon, Seth; Mungall, Chris</td>
</tr>
<tr>
<td>Archival bundle of GO data release.</td>
</tr>
</tbody>
</table>
Zenodo hosts 75% of world SW DOIs

Share ≠ Publish ≠ Preserve

Danger Zone

- Make this repository private
  - Please upgrade inveniosoftware-contrib
- Transfer ownership
  - Transfer this repository to another user or to an organization where you have the ability to create repositories.
- Archive this repository
  - Mark this repository as archived and read-only.
- Delete this repository
  - Once you delete a repository, there is no going back. Please be certain.

GitHub Guides

Making Your Code Citable

- Digital Object Identifiers (DOI) are the backbone of the academic reference and metrics system. If you’re a researcher writing software, this guide will show you how to make the work you share on GitHub citable by archiving one of your GitHub repositories and assigning a DOI with the data archive tool Zenodo.

PreTip: This tutorial is aimed at researchers who want to cite GitHub repositories in academic literature. Provided you’ve already set up a GitHub repository, this tutorial can be completed without installing any special software. If you haven’t yet created a project on GitHub, start first by uploading your work to a repository.
Thanks to Jose.Benito.Gonzalez@cern.ch, the CERN Digital Repository Section Leader, for providing valuable input for this presentation.