Enabling reproducible research: community practices, service needs and first lessons learnt

Dr Sünje Dallmeier-Tiessen
CERN

Reproducibility Workshop @TPDL
Hanover, September 2016
Agenda

Introduction

Terminology

Pragmatic approach

Perspectives (researcher, publisher, libraries, funder)

Use case: one research community

Service requirements

Challenges

CERN Open Data and Analysis Preservation

Lessons learnt
In order to reuse/repurpose results, you sometimes have to reproduce the original results first (to understand the exact details [1]). Conditions very discipline specific.
To reproduce or reuse research results a researcher needs…

- More than “just” the article
- Context, documentation
- Links to related research objects: data, code, workflows
- Understandable method, processing, software etc.
- Steps taken during the research process (versions)
A data publishing perspective: establishing context

Helicopter view: Trusted bridges across research life-cycle

Subsets of Data
Multiple Versions
Dynamic Data

Who?
When?
Where?

Linking data with data
Linking data with contributors
Linking data with articles
Linking data with institutions/funders

http://doi.org/10.5281/ZENODO.30799
http://doi.org/10.5281/ZENODO.30800
http://doi.org/10.5281/ZENODO.30800
http://doi.org/10.13039/501100000780

Slide credit to Trisha Cruse, Datacite
Our goal is to ensure that every researcher, at any phase of their career, or at any institution, will have seamless access to Persistent Identifiers (PIDs) for their research artefacts and their work will be uniquely attributed to them.

https://project-thor.eu/
Use Case: High-Energy Physics Community

Discussions, requirements and emerging services
CERN

Founded in 1954

Intergovernmental research organization

22 members states

~2500 employees

12,000 visiting scientists from over 70 countries and with 120 different nationalities

A different dimension of “collaborative research”

REPRODUCIBILITY WORKSHOP @TPDL 2016
A use case: High Energy Physics

- Small community, data driven
- Every experimental analysis with complex and big data and software pieces
- Experience with Open Access (it is the de facto default, in fact)
- Little or no experience with Open Science
- The usual: high throughput of personnel

There is only one LHC in the world: What does that mean for ..... reproducibility and replicability of an analysis?

- Surely it is work intensive, lots of dependencies
- What is needed? What makes sense for science?
First requirements

- Link articles and data/software, enable data discovery
- Incentivize open data and code sharing (data/software citation)

---

- Build further connections early in the research process towards a network of research objects
  - That enable collaborators to understand the research context
  - That can be searched for (internally) to accelerate research processes
  - Preservation
Impact

Open sourcing the secrets of the universe: huge amount of Large Hadron Collider data now online

By Sarah Kaplan  April 26

Cern makes 300TB of data available to download

By EMILY REYNOLDS  25 Apr 2016

Teilchenbeschleuniger LHC: 300 Terabyte Forschungsdaten freigegeben

heise online  26.04.2016  11:34 Uhr  -  Martin Holland
Measurement of the dependence of transverse energy production at large pseudorapidity on the hard-scattering kinematics of proton-proton collisions at $\sqrt{s} = 2.76$ TeV with ATLAS

Table 1

<table>
<thead>
<tr>
<th>cmenergies</th>
<th>observables</th>
<th>phrases</th>
<th>reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2760.0</td>
<td>SUMET, ET</td>
<td>Inclusive, Proton-Proton Scattering</td>
<td>P P → JET(S) X</td>
</tr>
</tbody>
</table>

Mean value of the sum of the transverse energy in $-4.9 < \eta < -3.2$ in pp collisions. Reported as a function of dijet $p_T$ and shown here for $+2.1 < \eta_{dijet} < +2.8$.

10.17182/hepdata.71318v1/t1

Table 2

Data from F2
10.17182/hepdata.71318v1
Mean value of the sum of the transverse energy in $-4.9 < \eta < -3.2$ in pp collisions, <SumET>. Reported...

Table 3

Data from F2
10.17182/hepdata.71318v1
Mean value of the sum of the transverse energy in $-4.9 < \eta < -3.2$ in pp collisions, <SumET>.

<table>
<thead>
<tr>
<th>R</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>RE</td>
<td>P P → JET(S) X</td>
</tr>
<tr>
<td>SQRT(S)</td>
<td>2760.0 GeV</td>
</tr>
<tr>
<td>YRAP</td>
<td>2.1-2.8</td>
</tr>
<tr>
<td>PTAVG [GEV]</td>
<td>SUMET [GEV]</td>
</tr>
<tr>
<td>50.10 - 63.10</td>
<td>10.2 ±0.1 stat ±0.1 syst</td>
</tr>
<tr>
<td>63.10 - 79.40</td>
<td>9.9 ±0.2 stat ±0.2 syst</td>
</tr>
</tbody>
</table>
Cranmer, Kyle S.

Personal Details (HepNames)

Name: Kyle S. Cranmer
Current Institution: New York U.
E-mail: cranmer@cern.ch
Links:
- http://theoryandpractice.org/
- https://www.linkedin.com/in/ky...
- http://twitter.com/KyleCranmer...
- https://github.com/cranmer

Fields: HEP-EX, HEP-PH, PHYSICS

Experiments: FNAL-E-0830, CERN-LHC-ATLAS, CERN-LEP-ALEPH

Identifiers: BA: K.S.Cranmer.1, INSPIRE: INSPIRE-00074922, ORCID: 0000-0002-5769-7094, ARXIV: cranmer_k_1

Publications

1. Data from figure 1 from: Search for gluinos in events with two same-sign leptons, jets and missing transverse momentum with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV
2. Data from figure 1 from: Search for gluinos in events with two same-sign leptons, jets and missing transverse momentum with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV
3. Additional data from: Search for gluinos in events with two same-sign leptons, jets and missing transverse momentum with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV
4. Data from figure 2 from: Search for gluinos in events with two same-sign leptons, jets and missing transverse momentum with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV
5. Data from figure 2 from: Search for gluinos in events with two same-sign leptons, jets and missing transverse momentum with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV
6. Data from figure 3 from: Search for gluinos in events with two same-sign leptons, jets and missing transverse momentum with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV

Citations Summary

747 papers found, 738 of them citeable (published or arXiv)

<table>
<thead>
<tr>
<th>Category</th>
<th>Citable papers</th>
<th>Published only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of papers analyzed:</td>
<td>738</td>
<td>638</td>
</tr>
<tr>
<td>Number of citations:</td>
<td>69629</td>
<td>66227</td>
</tr>
<tr>
<td>Citations per paper (average):</td>
<td>94.3</td>
<td>103.8</td>
</tr>
<tr>
<td>$\Phi$ index [?]</td>
<td>119</td>
<td>117</td>
</tr>
</tbody>
</table>

Breakdown of papers by citations:

<table>
<thead>
<tr>
<th>Category</th>
<th>Citable papers</th>
<th>Published only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renowned papers (500+)</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Famous papers (250-499)</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>Very well-known papers (100-249)</td>
<td>117</td>
<td>115</td>
</tr>
<tr>
<td>Well-known papers (50-99)</td>
<td>168</td>
<td>163</td>
</tr>
</tbody>
</table>
Barriers to practicing reproducible research

“We find that code, data, and ideas are each regarded differently in terms of how they are revealed and that guidance from scientific norms varies with pervasiveness of computation in the field.

The largest barriers to sharing are time involved in preparation of work and the legal Intellectual Property framework scientists face.” [6]
Moving upstream

In the research lifecycle
Considerations for service providers

Future purpose: reuse, reproducibility, preservation

What are the components of an analysis (where are they stored now)

How much do these components vary within the collaboration

How is quality defined

What are the dependencies (software, methods)

Versioning

Linking

Size (10-15TB per analysis)
A looooong form

Submission form with auto-complete functionality (based on connections made to existing databases within the collaboration)

WARNING: This is just a DEMO. Data saved is NOT backed-up at the moment and might be lost during any system upgrade.
Detailed physics metadata

Access via APIs to internal databases provides key information – CAP connects it.

Further information, such as OS, analysis software and related internal discussions, presentations and publications.

Detailed physics information (e.g. final state particles, cuts and vetos) for future reuse.
Reproducibility
1st lessons learnt

- Challenge of granularity, complexity, dependencies
- Solutions available to do data/software publishing, linking and data citation
- Applicable to other disciplines as well
- Moving upstream to enable reproducible research without “too much extra work”
- Role of docker, VMs?
Thanks to

CERN IT J. Delgado, J. Kunčar, T. Smith, T. Šimko
CERN SIS A. Dani, R. Dasler, P. Fokianos, P. Herterich, E. Maguire, A. Mattmann, L. Rueda
ALICE M. Gheata, M. Zimmermann
ATLAS K. Cranmer, L. Heinrich,
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
THOR Project
References


