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

Terminology

Repeatability

Replicability

Reproducibility

Conditions very discipline specific

Reusability

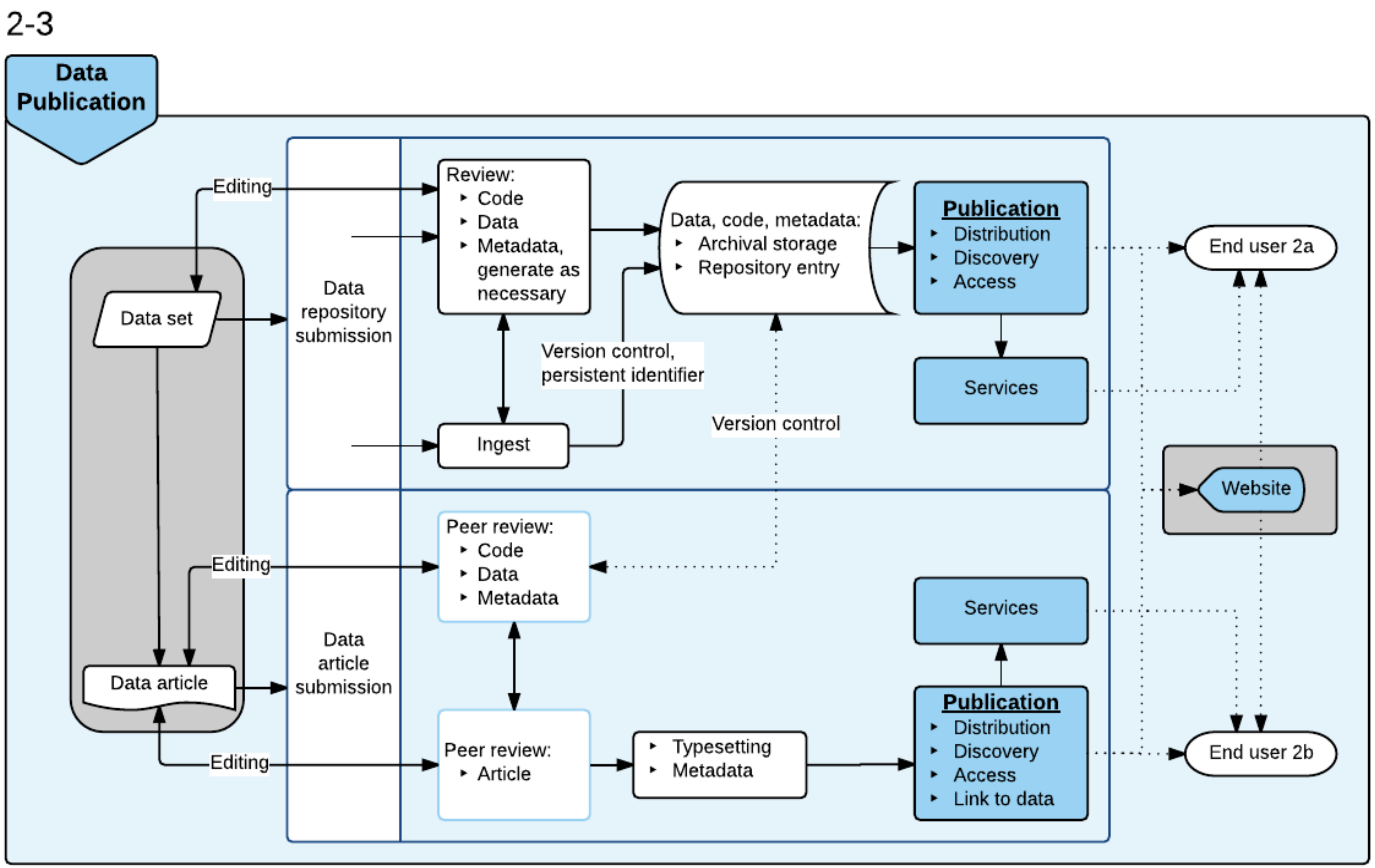
Repurposing

In order to reuse/repurpose results, you sometimes have to reproduce the original results first (to understand the exact details [1])

To reproduce or reuse research results a **researcher** needs…

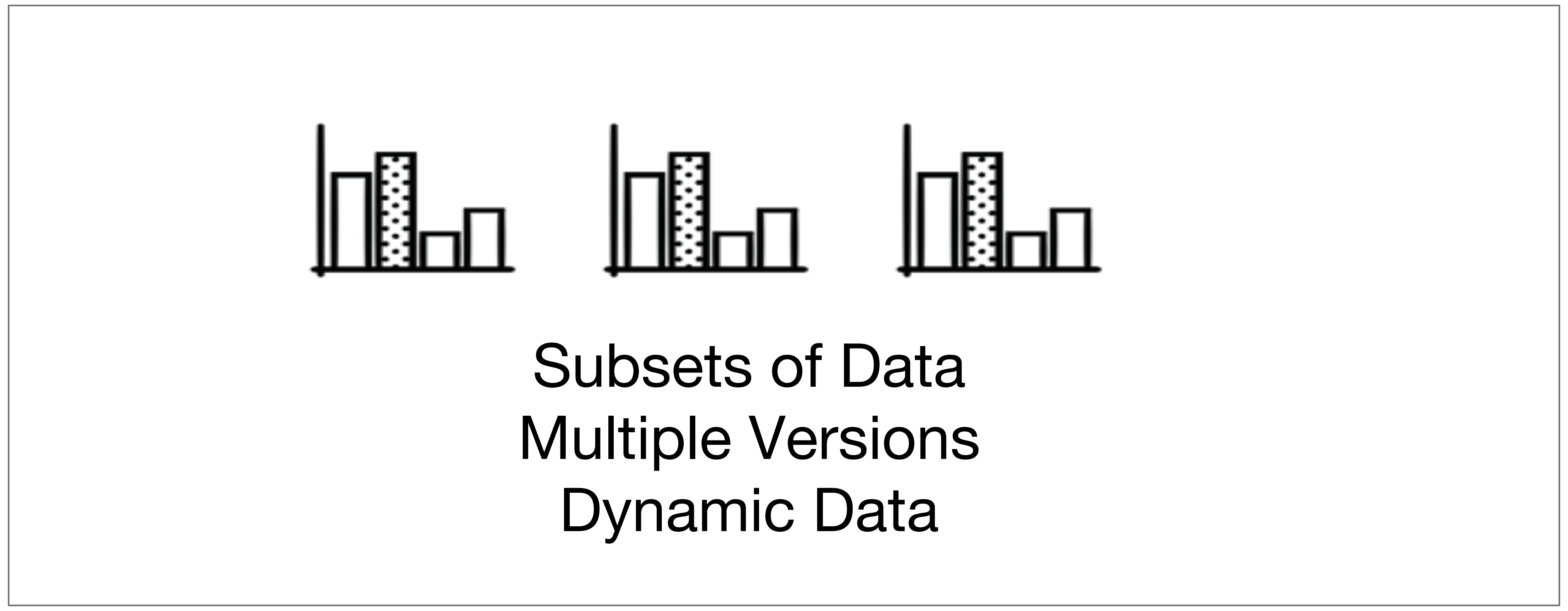
- More than “just” the article
- Context, documentation
- Links to related research objects: data, code, workflows
- Understandable method, processing, software etc.
- Steps taken during the research process (versions)

A data **publishing** perspective: establishing context

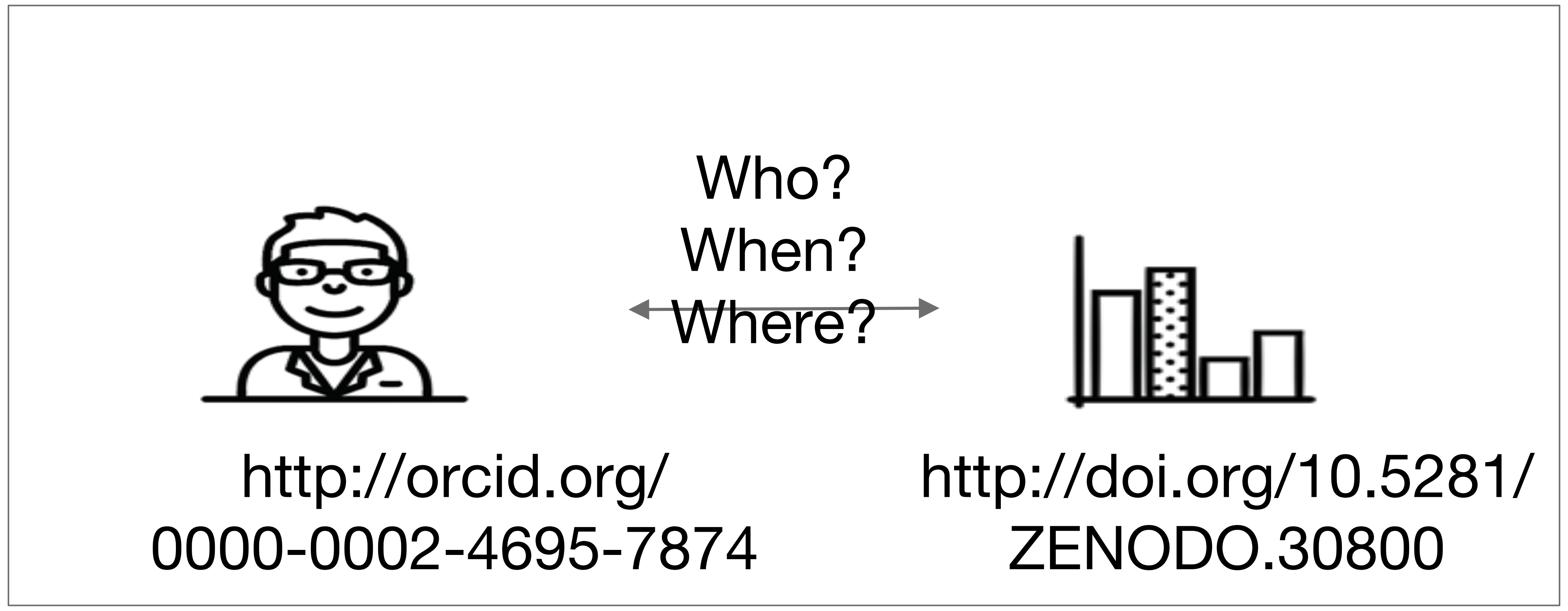


[2] DOI: 10.1007/s00799-016-0178-2

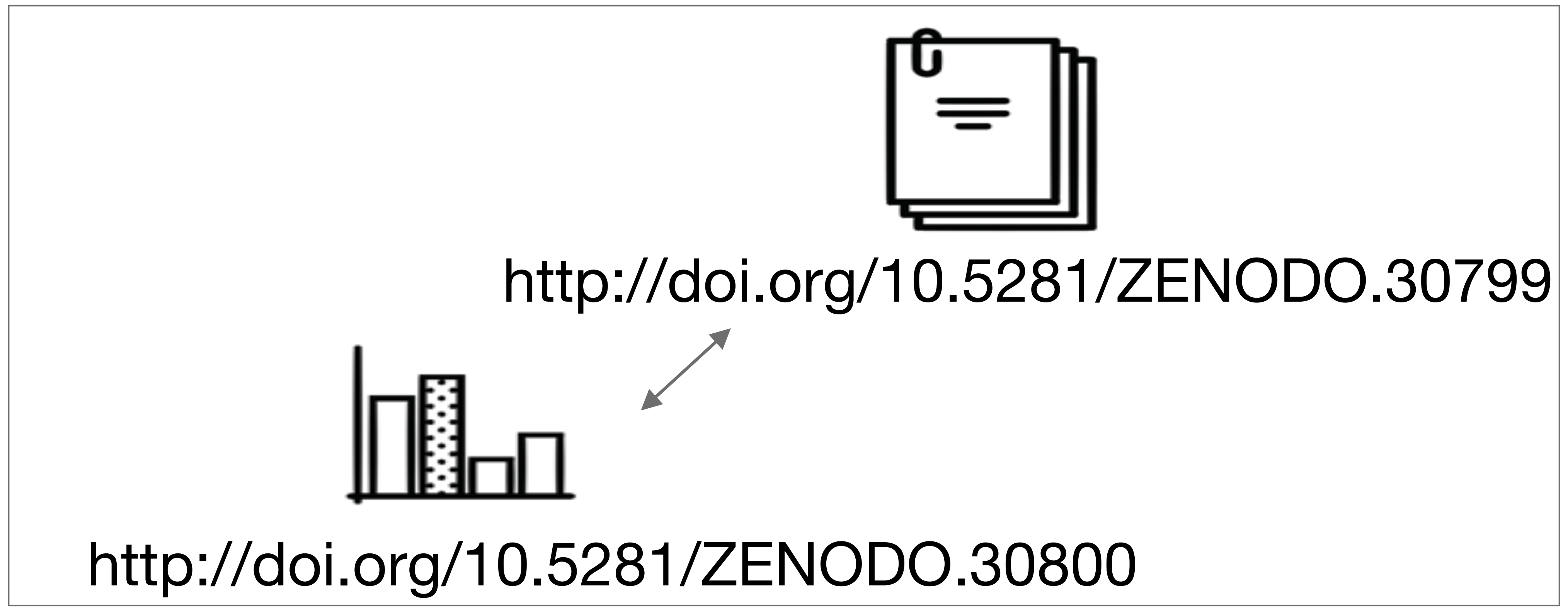
Helicopter view: Trusted bridges across research life-cycle



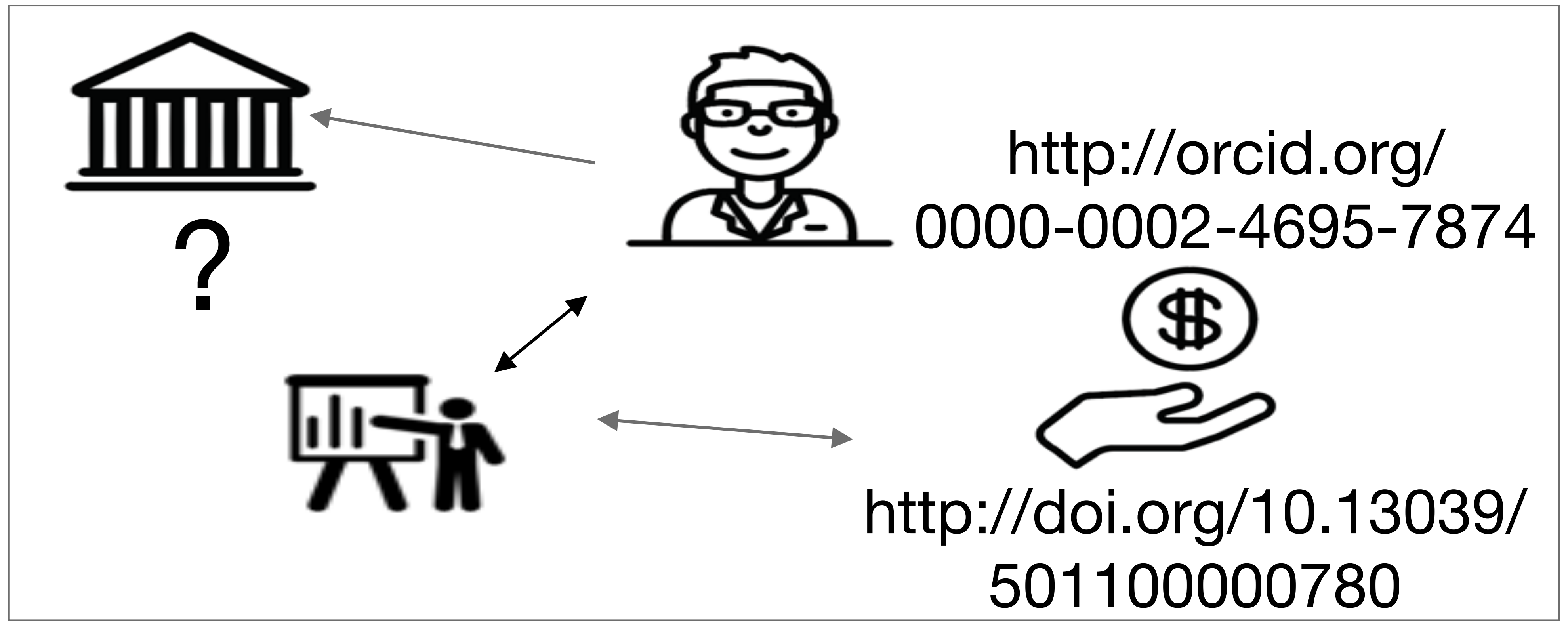
Linking data with data



Linking data with contributors



Linking data with articles



Linking data with institutions/funders

Slide credit to Trisha Cruse, Datacite

Technical and Human infrastructure for Open Research

https://project-thor.eu/

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

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"

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 ….

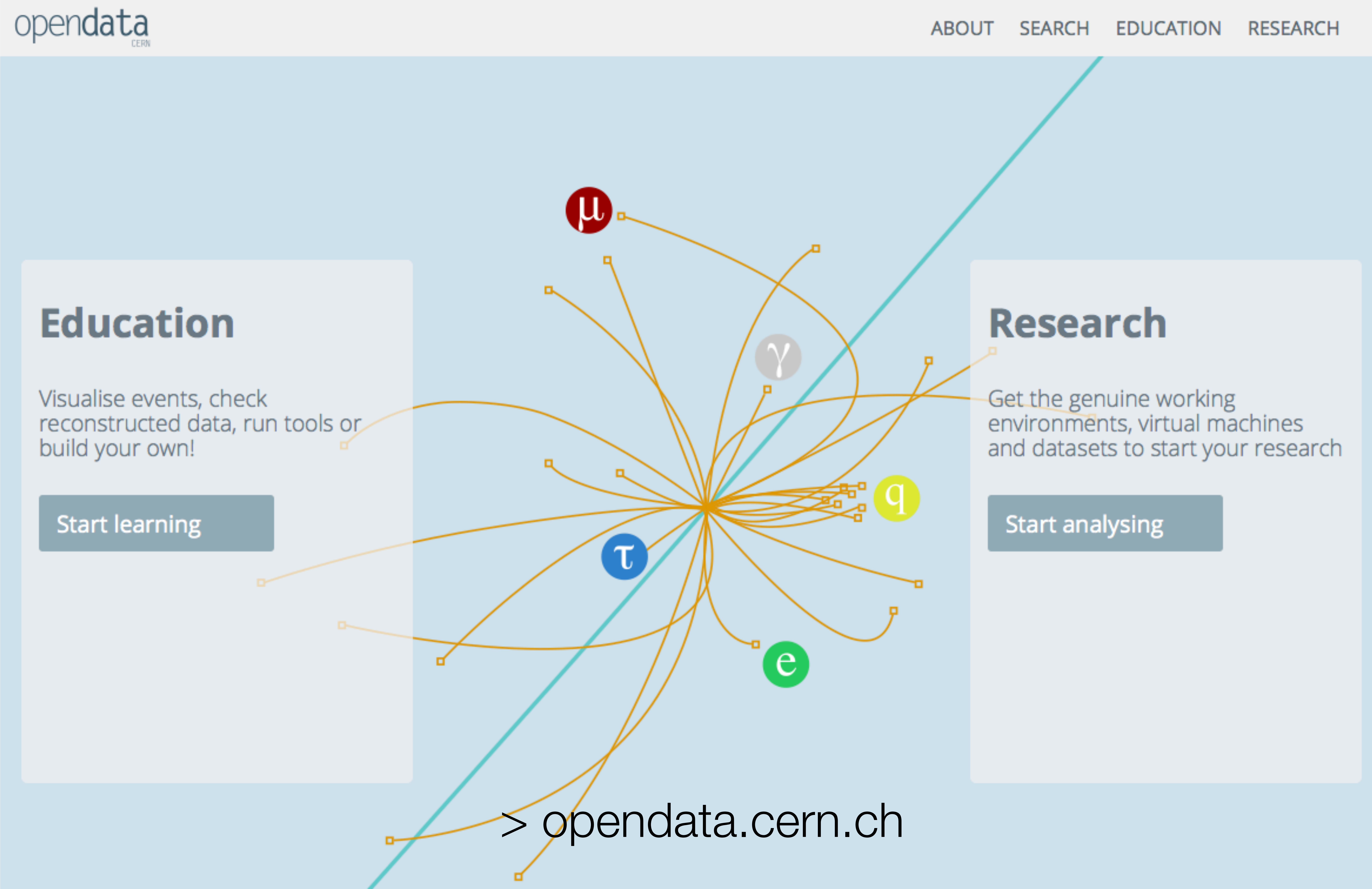
reproducibility and replicability of an analysis?

  - 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



> opendata.cern.ch

Impact

The Washington Post

Speaking of Science

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

By Sarah Kaplan April 26

WIRED SCIENCE

Science

Cern makes 300TB of data available to downloa

By EMILY REYNOLDS

25 Apr 2016

Teilchenbeschleuniger LHC: 300 Terabyte Forschungs freigegeben

heise online 26.04.2016 11:34 Uhr – Martin Holland

Hide Publication Information

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

Aad, Georges , Abbott, Brad , Abdallah, Jalal , Abdinov, Ovsat , Aben, Rosemarie , Abolins, Maris , AbouZeid, Ossama , Abramowicz, Halina , Abreu, Henso , Abreu, Ricardo
ATLAS

Phys.Lett. B756 (2016) 10-28, 2016

http://dx.doi.org/10.17182/hepdata.71318

DOI | View paper in Inspire | View old HepData

Abstract (data abstract)
CERN-LHC. The relationship between jet production in the central region and the underlying-event activity in a pseudorapidity-separated region is studied in 4.0 pb^{-1} of $\sqrt{s} = 2.76$ TeV pp collision data recorded with the ATLAS detector at the LHC. The underlying

Download All

Filter 34 data tabl

Table 1

http://www.hepdata.net/r JSON

Mean value of the sum of the transverse energy in -4.9 < eta < -3.2 in pp collisions, .
Reported as a function of dijet pT^avg, shown here for +2.1 < eta^dijet < +2.8.

10.17182/hepdata.71318.v1/t1

cmenergies
2760.0

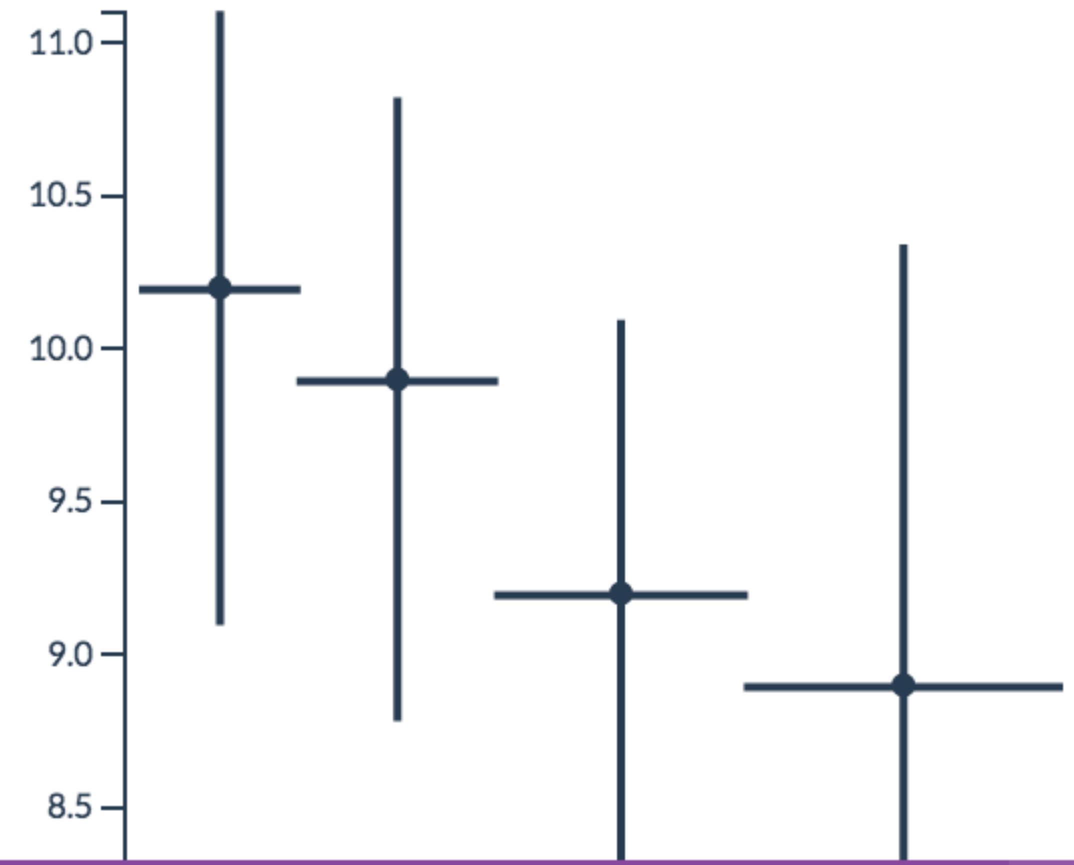
observables
SUMET
ET

phrases
Inclusive
Proton-Proton Scattering
Jet Production

reactions
P P --> JET(S) X

R	0.4
RE	P P --> JET(S) X
SQRT(S)	2760.0 GeV
YRAP	2.1-2.8
PTAVG [GEV]	SUMET [GEV]
50.10 - 63.10	10.2 ±0.1 stat +0.9 -1.1 sys
63.10 - 79.40	9.9 ±0.2 stat +0.9 -1.1 sys

Visualize



Cranmer, Kyle S.

PERSONAL INFORMATION

PUBLICATIONS AND OUTPUT

STATS

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

Period **Rank** **Institution**

Publications Datasets External

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

Co-Authors

B.Mellado.1 (13)
W.Quayle.1 (11)
C.T.Potter.1 (9)

Papers

	All papers	Single authored
All papers	747	11

Citations Summary

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

	Citeable papers	Published only
Number of papers analyzed:	738	638
Number of citations:	69629	66227
Citations per paper (average):	94.3	103.8
h_{HEP} index [?]	119	117

Breakdown of papers by citations:

	Citeable papers	Published only
Renowned papers (500+)	17	16
Famous papers (250-499)	21	20
Very well-known papers (100-249)	117	115
Well-known papers (50-99)	168	163

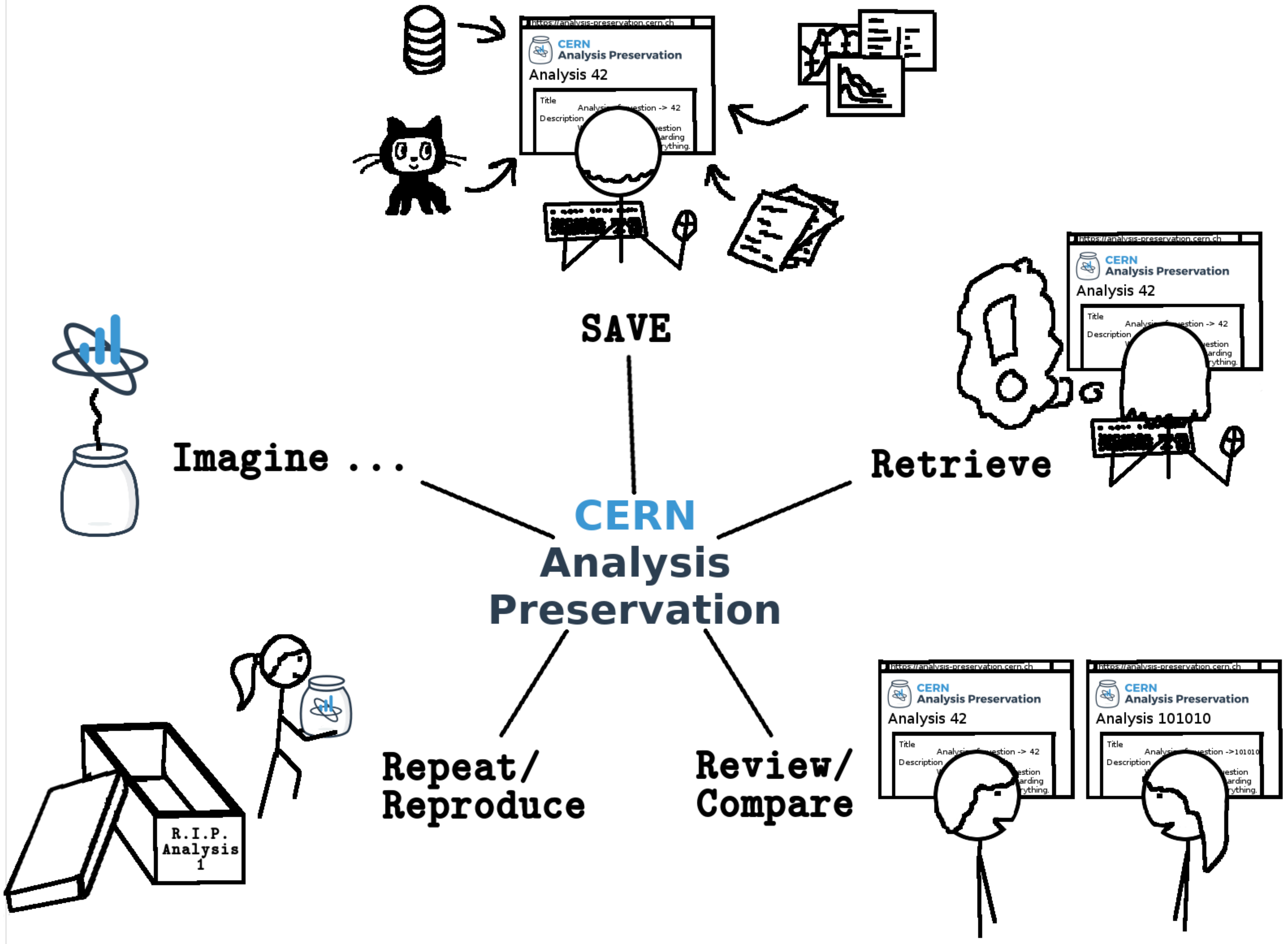
Barriers to practicing reproducible research

"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


CERN Analysis Preservation
SAVE
Imagine ...
Retrieve
Repeat/ Reproduce
Review/ Compare
R.I.P. Analysis 1
https://analysis-preservation.cern.ch
CERN Analysis Preservation
Analysis 42
Title
Description
Analysis 101010

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

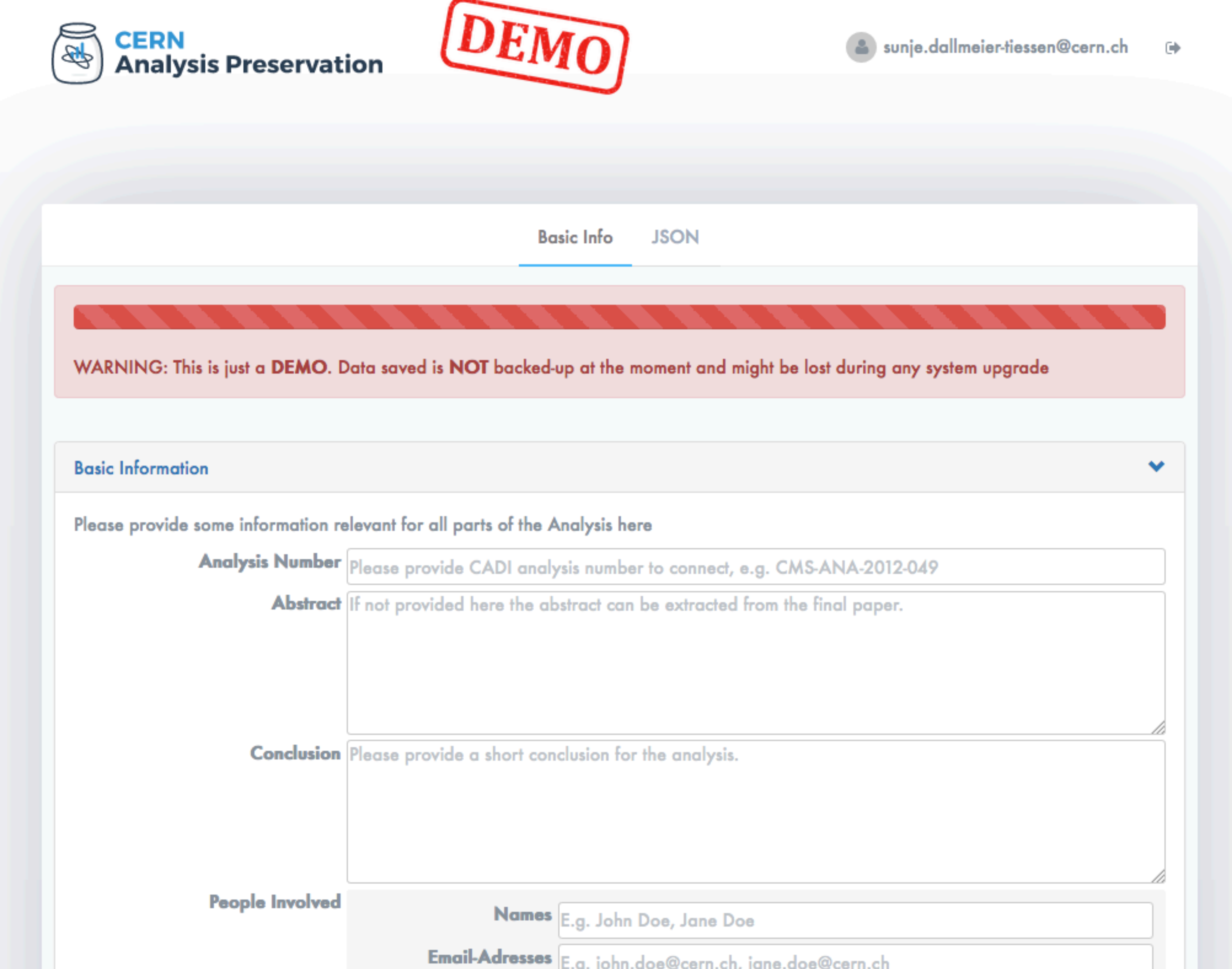
Linking

Size (10-15TB per analysis)


CERN
ANALYSIS PRESERVATION
DEMO
LOGIN
SIGN UP
Start typing
in
All Collections
9 records found.
ALICE
ALICE
ATLAS
ATLAS
CMS
CMS
LHCb
LHCb
© 2015 CERN Analysis Preservation
Privacy · Terms of Use · Cookies

A loooooong form

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

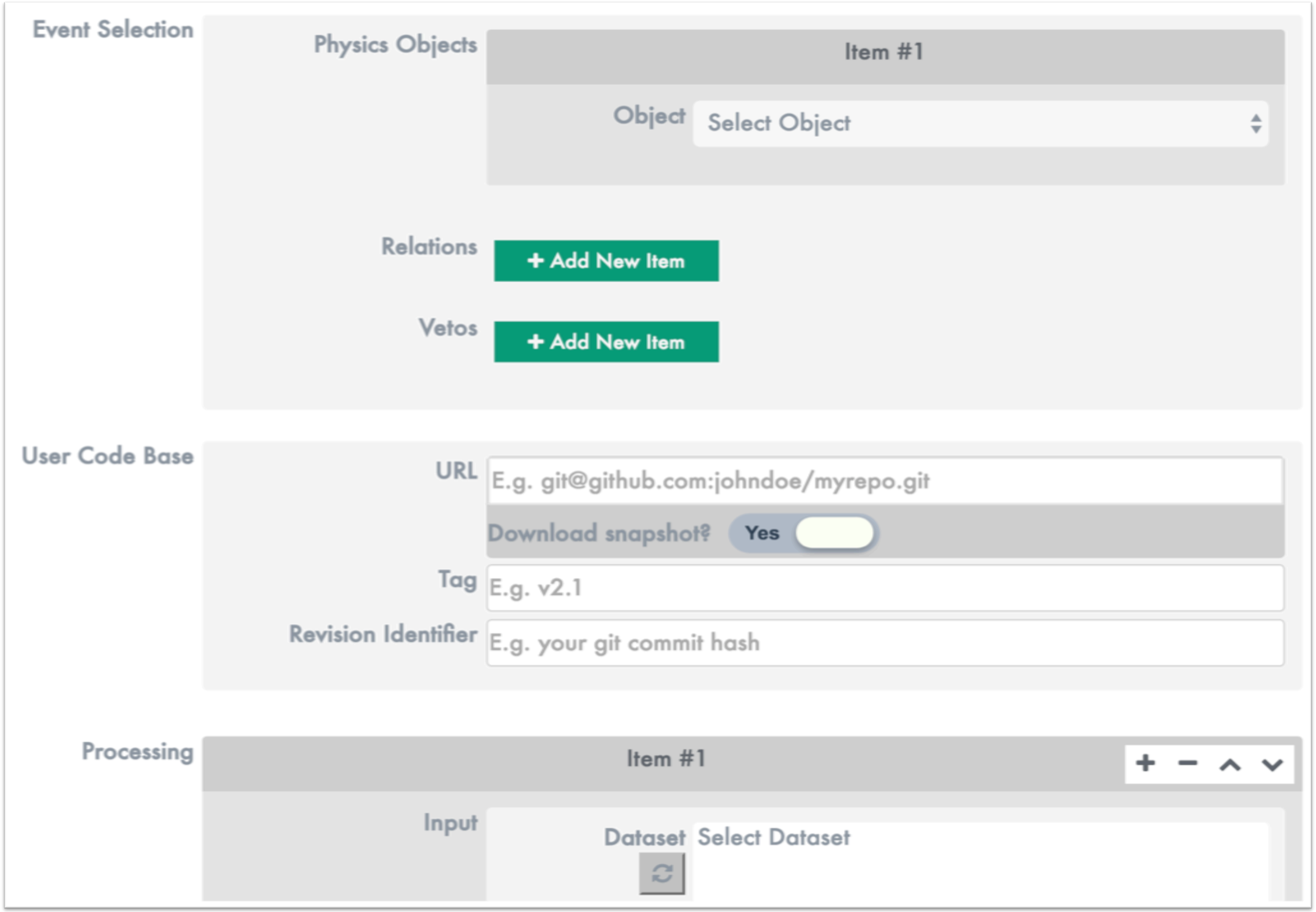


Detailed physics metadata

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

[1] http://web.stanford.edu/~vcs/papers/TrustYourScience-STODDEN.pdf

[2]Key Components in Data Publishing: DOI: 10.1007/s00799-016-0178-2

[3] http://opendata.cern.ch/

[4] www.hepdata.net

[5] www.inspirehep.net

[6] Stodden, Victoria, The Scientific Method in Practice: Reproducibility in the Computational Sciences (February 9, 2010). MIT Sloan Research Paper No. 4773-10. Available at SSRN: http://ssrn.com/abstract=1550193