Translating Open Science ideals to actions for scientists

Xiaoli Chen[1], **Sebastian Feger**[2], Robin Dasler and Sünje Dallmeier-Tiessen

CERN
and
[1]University of Sheffield,
[2]Universität Stuttgart


LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
CMS
ALICE
LHC 27 km

CERN

CERN

Intergovernmental research organization

22 member states

2200 employees but 10000 users on site

70 countries, 120 different nationalities

A different dimension of collaborative research

Open Science Ideals

NATURE | EDITORIAL

Empty rhetoric over data sharing slows science

Governments, funders and scientific communities must move beyond lip-service and commit to data-sharing practices and platforms.

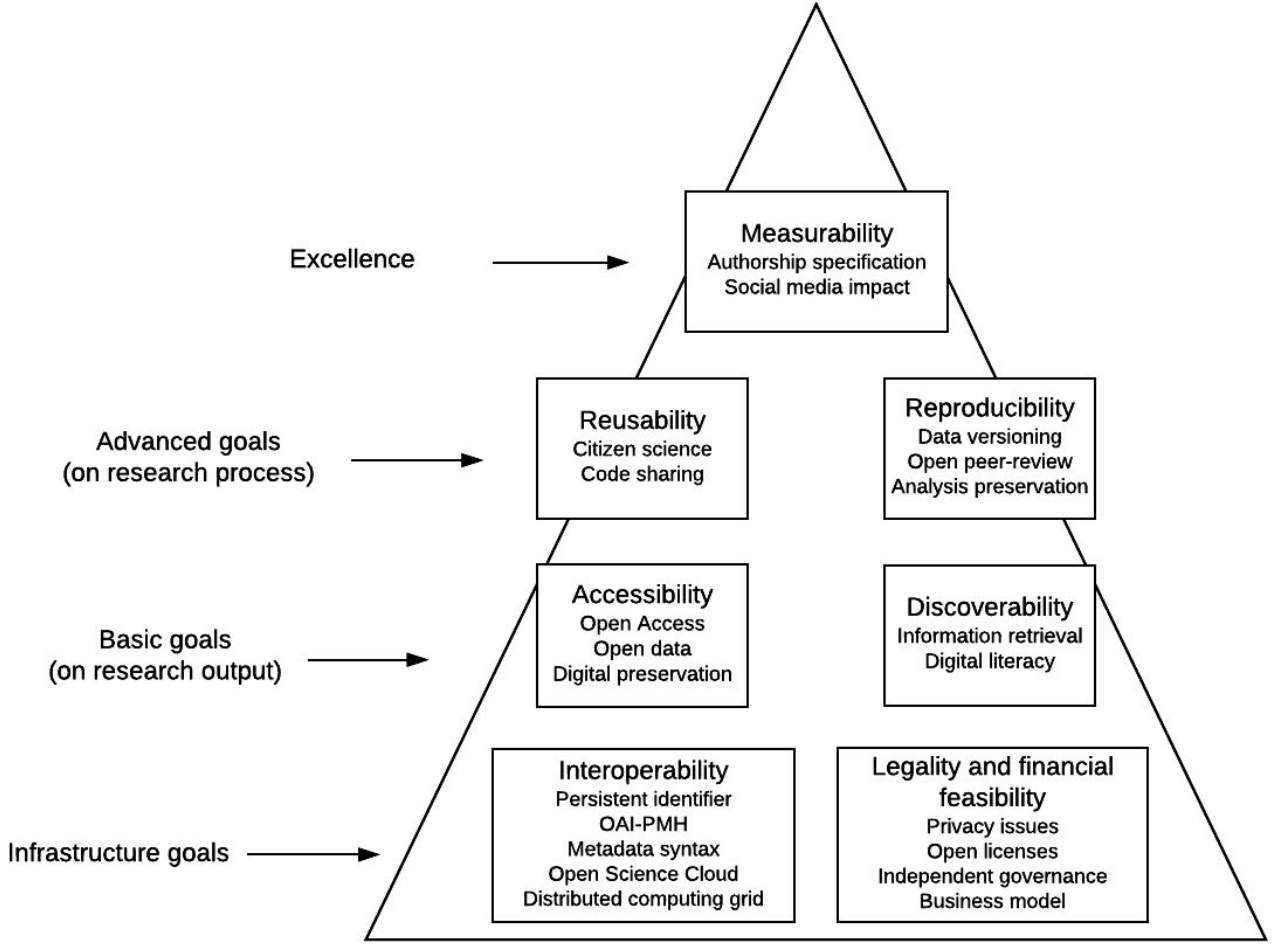
12 June 2017

New territory

New requirements

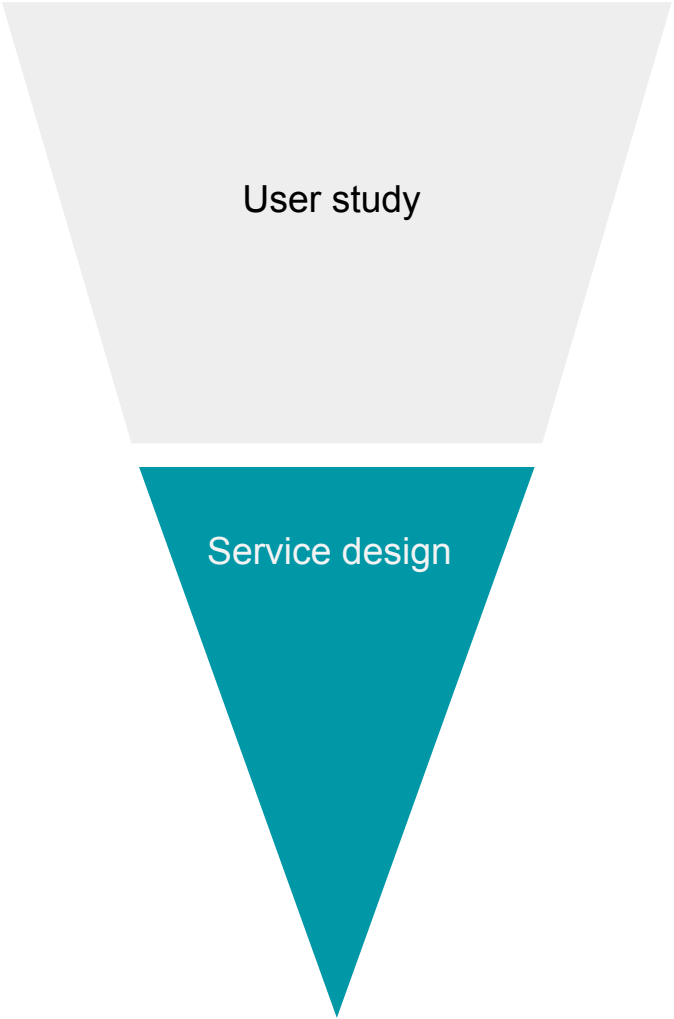
New opportunities

New challenges



(Chen et al., work in progress)

User-centered design



User-centered design in science *(d)*

- “flipping the perspective from a technical to a human-centered approach changes the perceived benefits and design goals” *(a)*

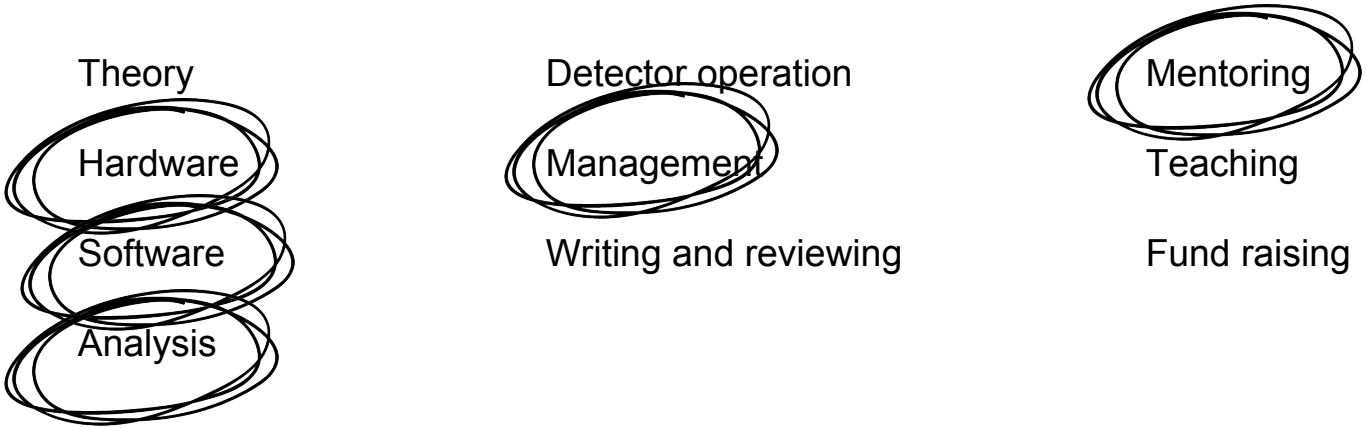
- Even small interface changes of analysis systems impact scientist’s behaviors *(b)*

Approach: interview based, with a representative set of researchers, junior to senior, different experiments

Drivers: Open Science in HEP

Findings - array of responsibilities

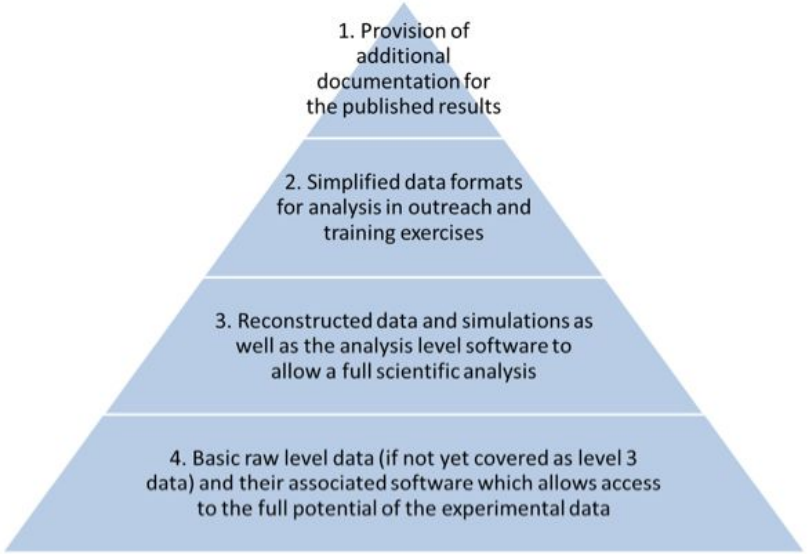
A physicist can simultaneously be a researcher, a collaborator and an academic



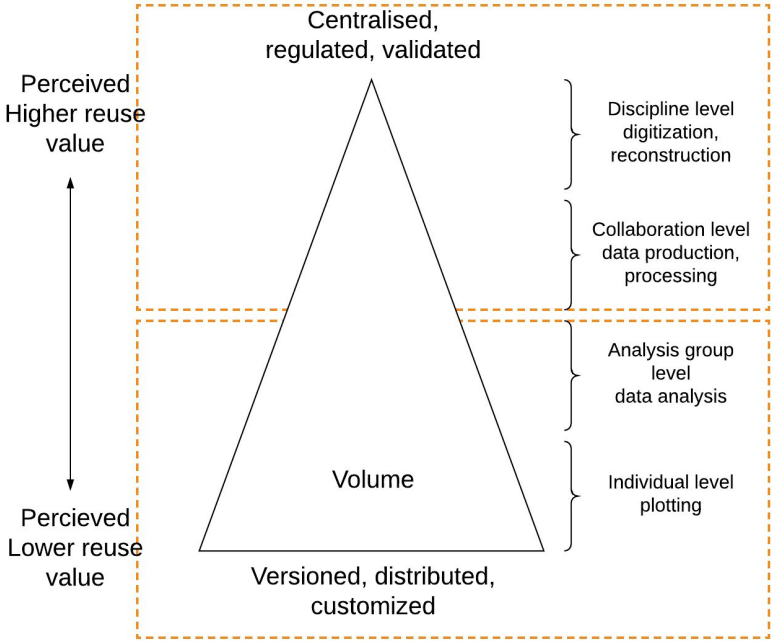
HEP research is demanding, priorities are highly selective

Information services operate **within and among** each block of work

Findings - data and software



(DPHEP Study Group, 2009)



(Chen et al., work in progress)

In terms of information service for data and software, different type of digital objects need to be accommodated separately

Findings - perception on sharing and documenting

- Those required to share: ensure intelligibility, on distributed platforms, depend on human network for discoverability
- Sharing and documenting decision based primarily on perceived relevance, quality, reuse value, potential impact
- Unclear sense of ownership of research object

Certain levels of openness serves certain degrees of reproducibility and reusability

Discussions

How far away are the physicists from science in the open?

For scientists, open practices are documenting and sharing, and the purpose is reproducibility and reuse.

How to make sharing and documenting “a good deal” for the scientists?

Mainly for the experimentalists, to whom the traditional metrics and incentives barely means anything

Harmonizing the Open Science goal for scholarly communication and researchers community

Translating Open Science ideals to actions for scientists

… the actions

How can we build a service to foster reproducible research? Is that possible?

Preserving a physics analysis

Minimize the burden

- Meet the “normal research flow”: submission, updating through terminal/shell
- Submission form designed to support documentation
  - Tailored to collaborations
  - Autosuggest and autocomplete

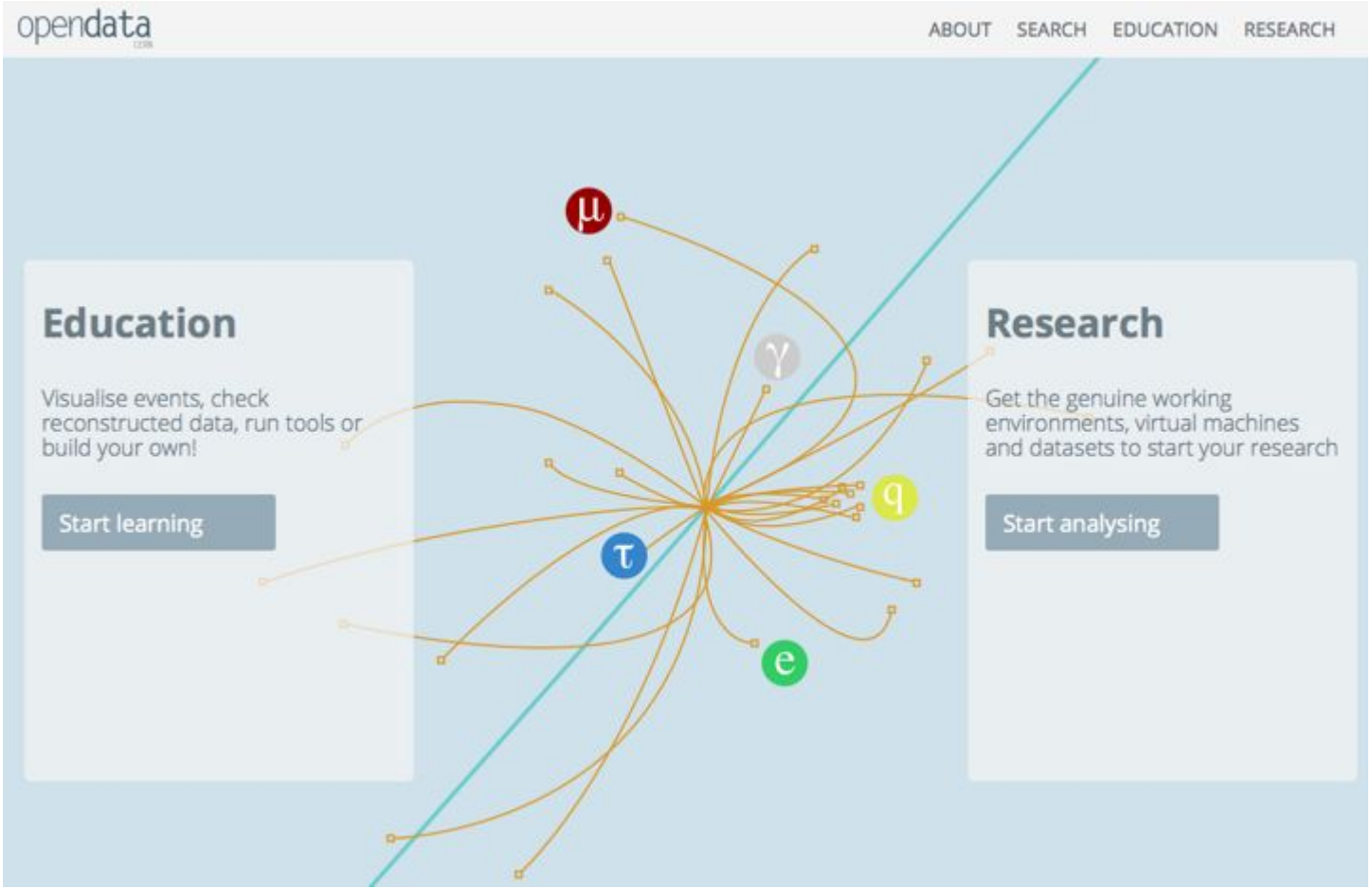
CERN
Analysis Preservation

Preserving a physics analysis - More than just knowledge documentation

- Provide documenting scientists with **tangible** benefits
  - Benefits that affect day-to-day analysis work, e.g. findability
  - Reproducibility is a high-level, long-term goal that does not always play well as a motivating argument at the moment - visualize possible impact
- Opportunity to foster **collaboration** (based on increased visibility)
  - Who does what, who uses what
  - Who can help with...
- Structured submission forms act as templates
  - Comprehensive documentation made easy - spot missing pieces
  - Discover issues / conflicts in the analysis workflow early

Set of incentives that encourage documentation and sharing of **ongoing analyses**

Opening up



http://opendata.cern.ch/

The Washington Post

Speaking of Science

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

By Sarah Kaplan April 26

WIRED SCIENCE

Science

Cern makes 300TB of data available to download

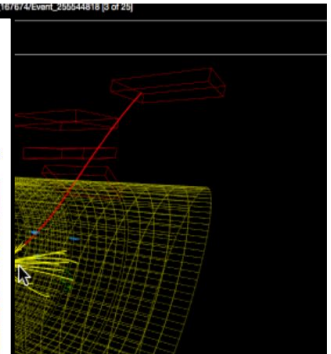
By **EMILY REYNOLDS**

25 Apr 2016

Teilchenbeschleuniger LHC: 300 Terabyte Forschungsdaten freigegeben

heise online 26.04.2016 11:34 Uhr – Martin Holland

vorlesen



arXiv:1704.05842

High Energy Physics - Phenomenology

Jet Substructure Studies with CMS Open Data

Aashish Tripathee, Wei Xue, Andrew Larkoski, Simone Marzani, Jesse Thaler

(Submitted on 19 Apr 2017 (v1), last revised 8 May 2017 (this version, v2))

We use public data from the CMS experiment to study the 2-prong substructure of jets. The CMS Open Data is based on 31.8/pb of 7 TeV proton-proton collisions recorded at the Large Hadron Collider in 2010, yielding a sample of 768,687 events containing a high-quality central jet with transverse momentum larger than 85 GeV. Using CMS's particle flow reconstruction algorithm to obtain jet constituents, we extract the 2-prong substructure of the leading jet using soft drop declustering. We find good agreement between results obtained from the CMS Open Data and those obtained from parton shower generators, and we also compare to analytic jet substructure calculations performed to modified leading-logarithmic accuracy. Although the 2010 CMS Open Data does not include simulated data to help estimate systematic uncertainties, we use track-only observables to validate these substructure studies.

Comments:	35 pages, 19 figures, 6 tables, source contains sample event and additional plots; v2: references updated and figure formatting improved
Subjects:	**High Energy Physics - Phenomenology (hep-ph)**; High Energy Physics - Experiment (hep-ex)
Report number:	MIT-CTP 4890
Cite as:	arXiv:1704.05842 **[hep-ph]** (or arXiv:1704.05842v2 **[hep-ph]** for this version)

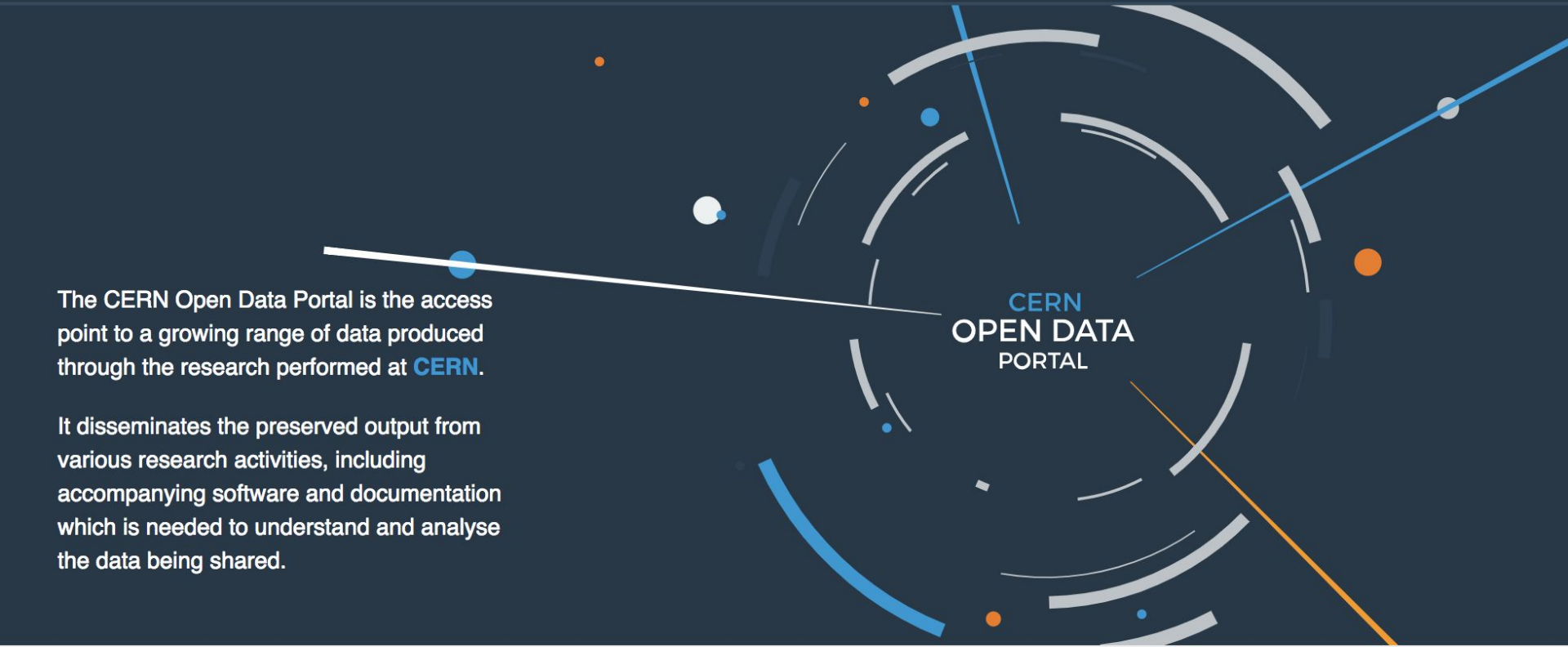
Submission history

From: Jesse Thaler [view email]
[v1] Wed, 19 Apr 2017 18:00:00 GMT (28272kb,AD)
[v2] Mon, 8 May 2017 01:19:34 GMT (25903kb,AD)

opendata CERN

Search

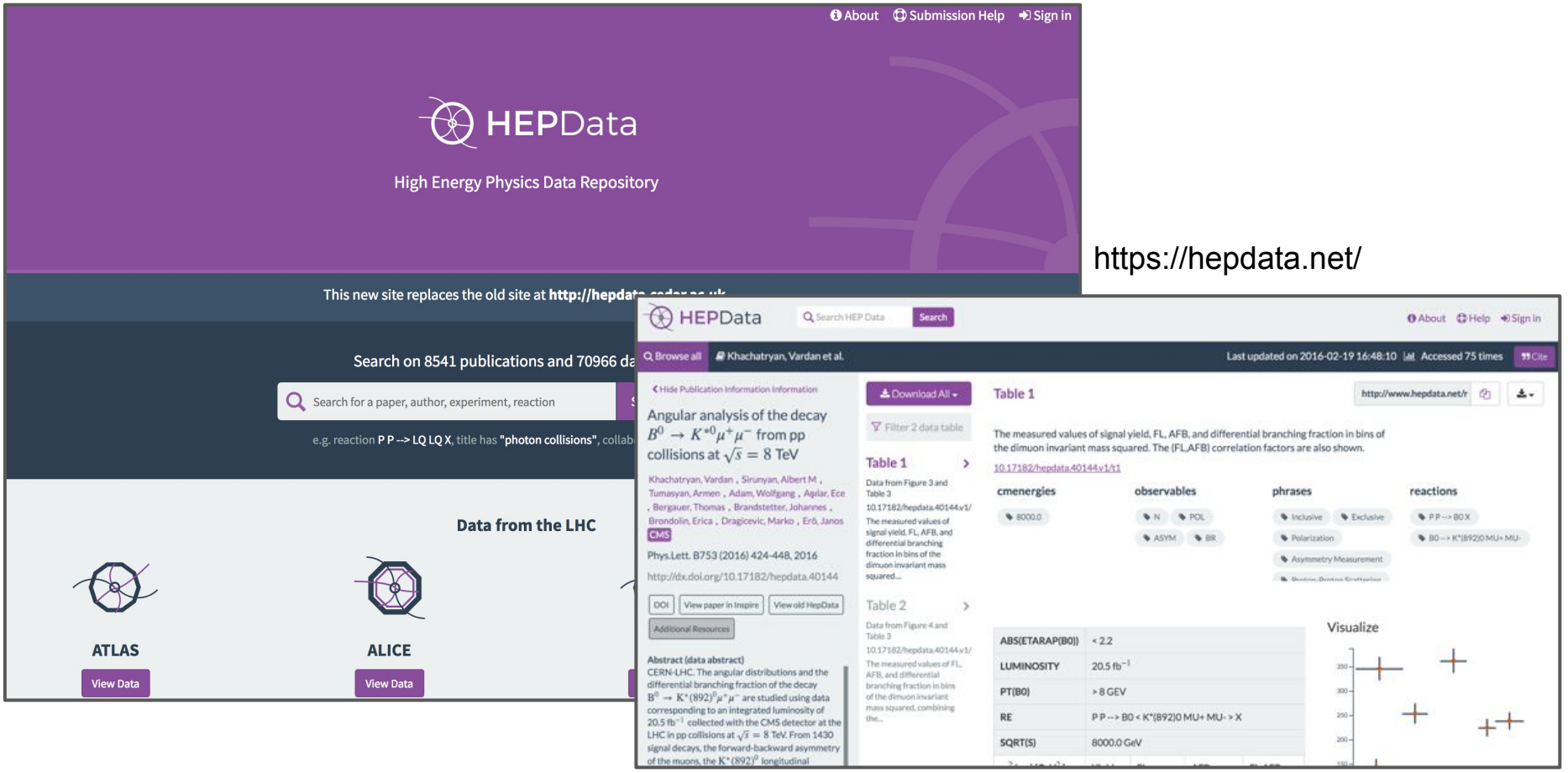
About



The CERN Open Data Portal is the access point to a growing range of data produced through the research performed at CERN.

It disseminates the preserved output from various research activities, including accompanying software and documentation which is needed to understand and analyse the data being shared.

Opening “things” up



https://hepdata.net/

Conclusions

We want to build services that respond to the community's growing demands

How do we do this best - so that it works?

We have to work together with the community

- Continuously consult and test with the physicists/users
- Incorporate expert insight into service design - generic functions only go so far
- Build services that are integral to the research workflow
- Create incentives that matter to the physicists

Implications for Open Science and reproducible research:

- Invite the researchers community to Open Science by stressing reproducibility and reusability
- Build off the rich content and knowledge offered by the community and extend their impact

References / Literature

(a) Jesper Molin, Paweł W. Woźniak, Claes Lundström, Darren Treanor, and Morten Fjeld. 2016. Understanding Design for Automated Image Analysis in Digital Pathology. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (NordiCHI '16). ACM, New York, NY, USA, Article 58, 10 pages. DOI: https://doi.org/10.1145/2971485.2971561

(b) Radu Jianu and David Laidlaw. 2012. An evaluation of how small user interface changes can improve scientists' analytic strategies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 2953-2962. DOI: http://dx.doi.org/10.1145/2207676.2208704

(c) Norman, D. A., & Draper, S. W. (Eds.). (1986). User centered system design: new perspectives on human-computer interaction. Hillsdale, N.J: L. Erlbaum Associates.

(d) REANA Github Page. https://github.com/reanahub

(e) FORCE11. 2014. The FAIR data principles. Website. (2014). Retrieved August 8, 2017 from https: //www.force11.org/group/fairgroup/fairprinciples.