for
Medical Research and HealthCare

Tim Smith CERN /IT

at CERN OpenLab “IT in HealthCare” Workshop

Data is the Answer…

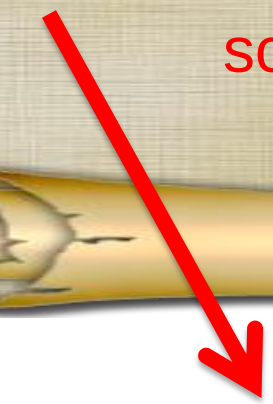
Digital Dark Ages

Scientific method

Propose hypotheses to explain phenomena

Test hypotheses predictions through repeatable experiment

Share observations and conclusions for independent scrutiny, reproduction and verification

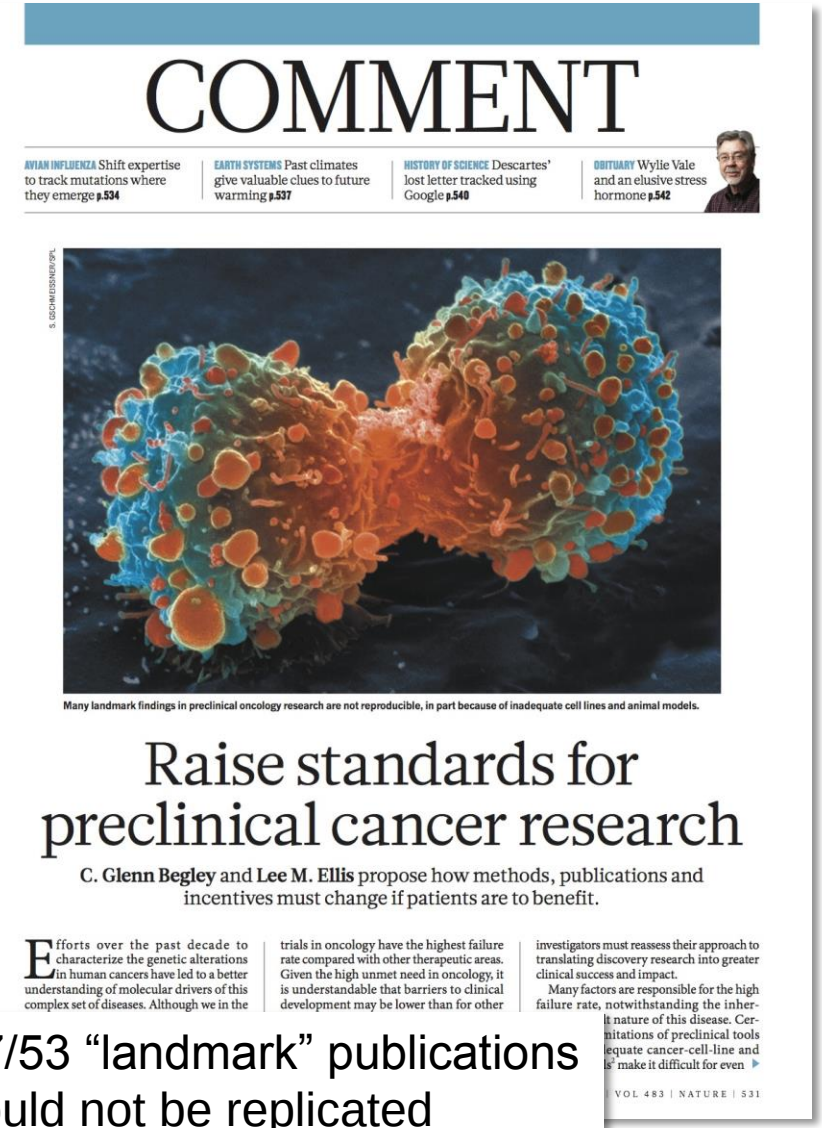


Publication: Preparation (standardization), issuing

Unavailable / Unverified

Error prone

Biologists must realize the pitfalls of work on massive amounts of data.



COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534 | EARTH SYSTEMS Past climates give valuable clues to future warming p.537 | HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540 | OBITUARY Wylie Vale and an elusive stress hormone p.542

S. GSCHMEISSNER/SPL

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inher-

47/53 "landmark" publications could not be replicated

Policies: Open and Managed

the WHITE HOUSE PRESIDENT BARACK OBAMA

BRIEFING ROOM | ISSUES | THE ADMINISTRATION | PARTICIPATE

Home • Briefing Room • Presidential Actions • Executive Orders

The White House

Office of the Press Secretary

E-Mail | Tweet | Share

For Immediate Release

May 09, 2013

Executive Order -- Making Open and Machine Readable the New Default for Government Information

EXECUTIVE ORDER

Guidelines on Open Access
to Scientific Publications and Research Data
in Horizon 2020

RESEARCH COUNCILS UK

Accessibility | Media Enquiries

Home | Funding | Research | Innovation | Skills | Public Engagen

Home / Research / RCUK Common Principles on Data Policy

RCUK Common Principles on Data Policy

Making research data available to users is a core part of the Research Councils' ren
to transparency and to a coherent approach across the research base. These RCUK
framework for individual Research Council policies on data policy.

Principles

- Publicly funded research data are a public good, produced in the public interest
restrictions as possible in a timely and responsible manner that does not harm i
- Institutional and project specific data management policies and plans should be
practice. Data with acknowledged long-term value should be preserved and ren

NIH U.S. National Library of Medicine

Databases | Find, Read, Learn | Exp

NIH Trans-NIH BioMedical Info

Home

NIH Data Sharing Policies

This table lists data sharing policies in effec
investigators and data. Individual requests
expectations for data sharing that apply to

Show 50 entries

IC	Data Sharing Policy Name	Description of Data Sharing Policy
NHGRI	ENCODE	Requires resource producers to release primary data along with an

wellcome trust

Our vision | Funding | Managing a grant | Education resources

Publications | Jobs | Policy | Strategy | Organisation | History | Timeline | Logo usage | Contac

Policy and position statements

Consultation responses

Policy on data management and sharing

The Wellcome Trust is committed to ensuring that the outputs of the rese
and used in ways that maximise public benefit. Making research data wi
and responsible manner ensures that these data can be verified, built up
to generate improvements in health.

Make your data Available

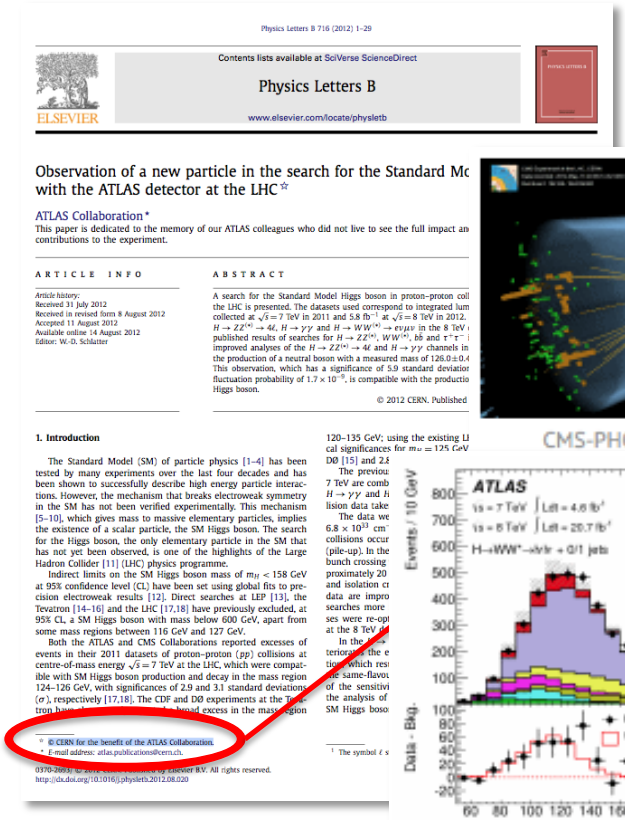
Google

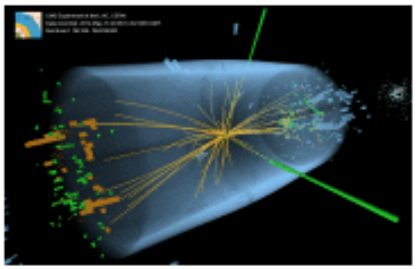
404. That's an error.

The requested URL /research_data was not found on this server. That's all we know.

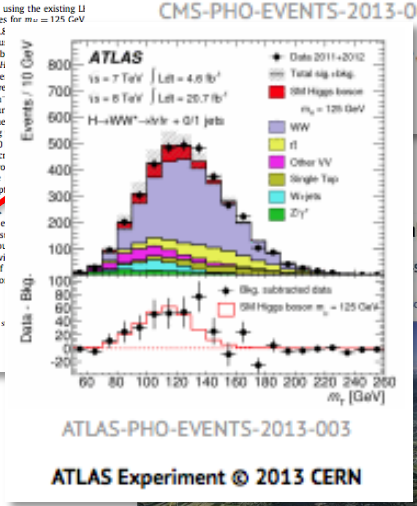
- 2008: MEDLINE study: URL decay 20%
- 2011: arXiv: 28% referenced URLs don't exist
  - 45% exist but are not archived
- 2013: IJEDICT study of Theses: 58% URLs broken

Digital Libraries



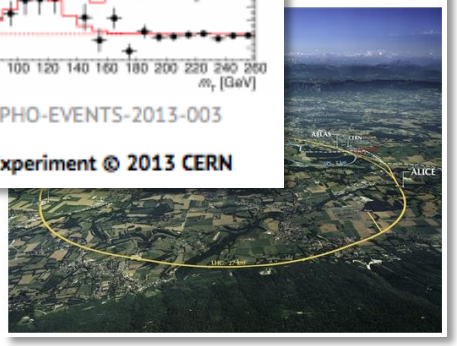


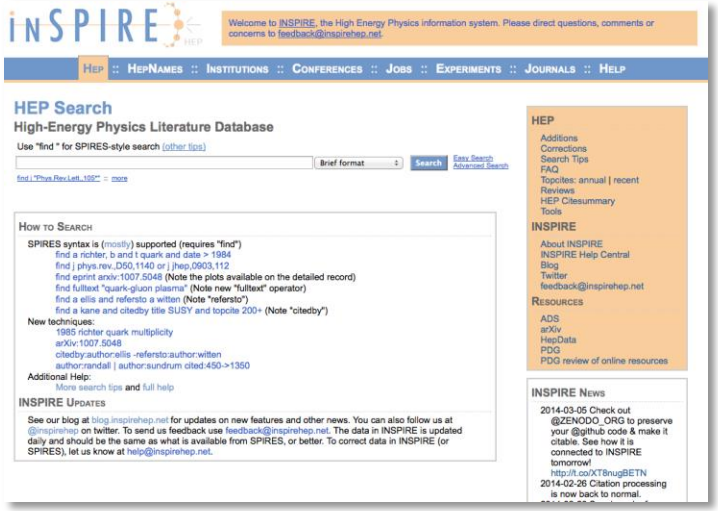
CMS-PHO-EVENTS-2013-003



ATLAS-PHO-EVENTS-2013-003

ATLAS Experiment © 2013 CERN



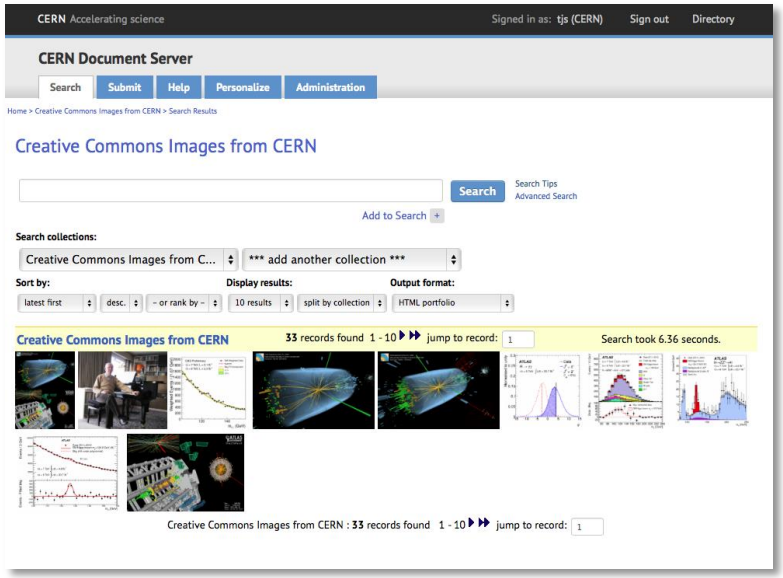


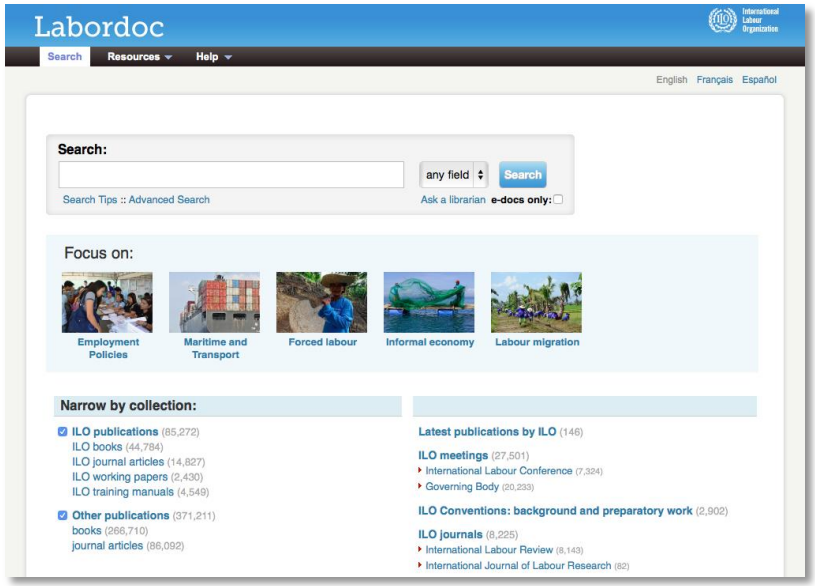
Open Source Repository Platform

INVENIO

- Mature **digital library** platform, originated in 2002 at CERN
  - **OAIS**-inspired **preservation** practices
- Co-developed by international collaboration

CERN · DESY · EPFL École Polytechnique Fédérale de Lausanne · Fermilab · SLAC National Accelerator Laboratory · International Labour Organization · Jülich · GSI · UAB Universitat Autònoma de Barcelona · Universidad Zaragoza · réro





Big Data … *in small pieces*

Big facilities
x (a small number)

Dedicated Big Data Stores

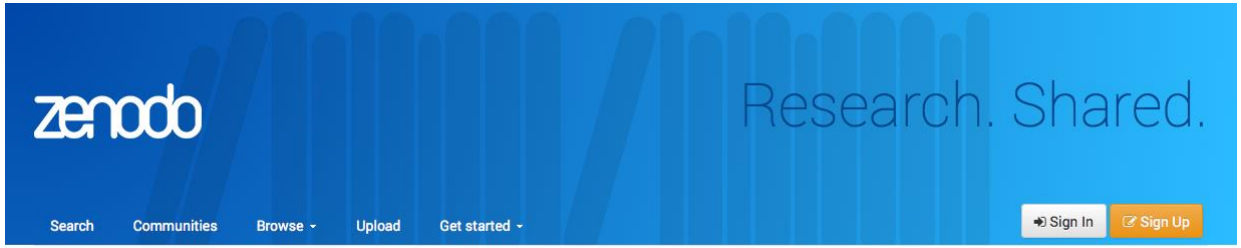
+ INVENIO

Data Size

zenodo

Long tail of science
x (a large number)

Open Data as a service



18 June 2014

Dataset | Open access

Drosophila simulans template brains

Ostrovsky, Aaron D. ; Goetz, Lea ; Jefferis, Gregory S. X. E.

(show affiliations)

Male and female symmetric averaged templates (11 and 10 brains, respectively) and intersex template brain for *Drosophila simulans*. Voxel size: (0.461, 0.461, 1) micron.

Files

Name	Date	Size	
DsimF.nrrd	20 Jun 2014	43.3 MB	Download
DsimIS.nrrd	20 Jun 2014	48.2 MB	Download
DsimM.nrrd	20 Jun 2014	115.4 MB	Download

Comments

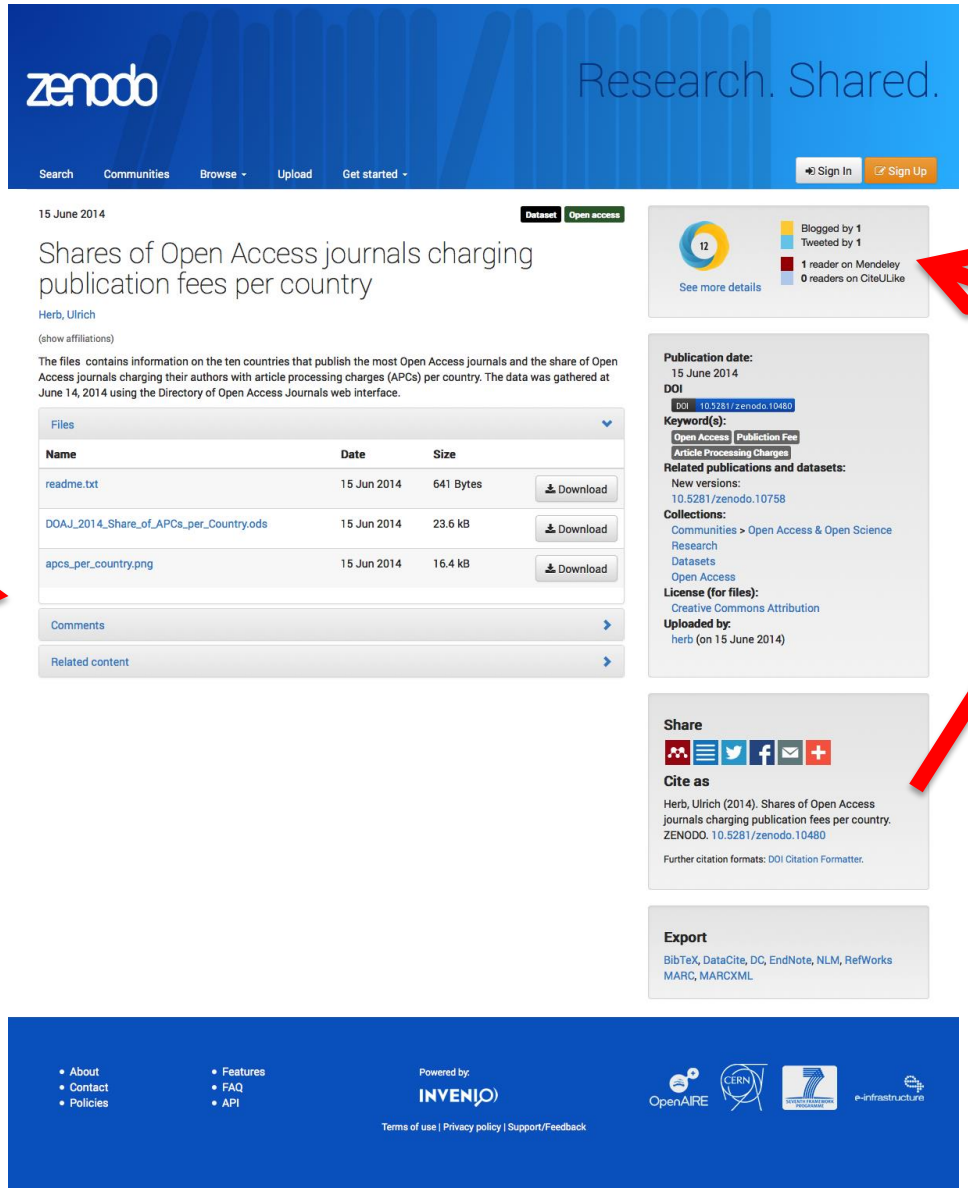
Related content

Publication date:
18 June 2014
DOI
DOI 10.5281/zenodo.10594
Keyword(s):
drosophila | neuroanatomy
Grants:
OLFPERCEPT - Neural Basis of Olfactory Perception
Collections:
Communities > European Commission Funded Research (Open AIRE)
Communities > Drosophila Datasets
Open Access
License (for files):
Creative Commons CCZero
Uploaded on:
19 June 2014

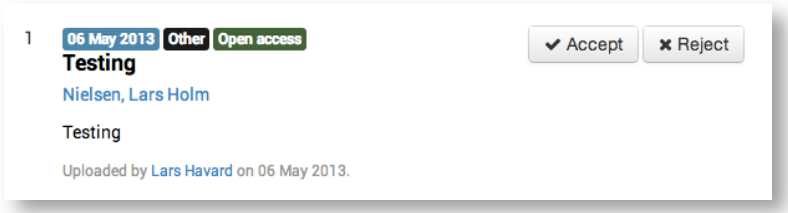
New to ZENODO?
Read more about features and benefits.
Sign Up

```
<contributor contributorType="Funder">
  <contributorName>European Commission</contributorName>
  <nameIdentifier nameIdentifierScheme="info">
     info:eu-repo/grantAgreement/EC/FP7/282896
  </nameIdentifier>
</contributor>
```

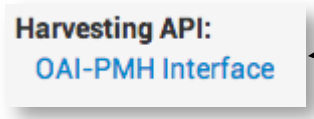
Low Barriers



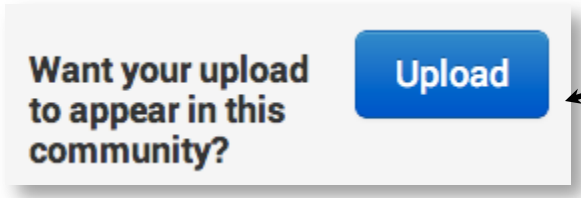
Communities



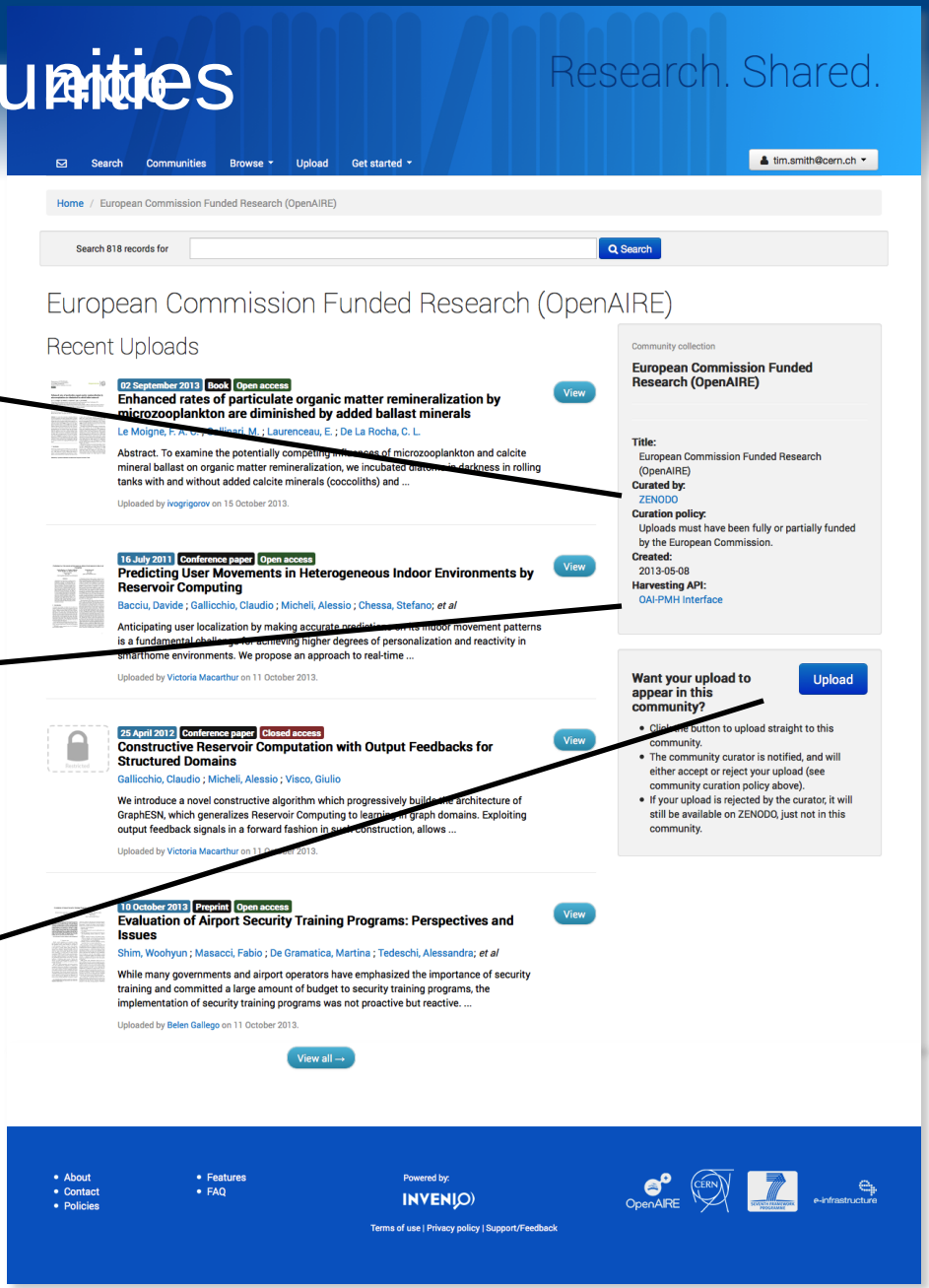
Accept/reject uploads



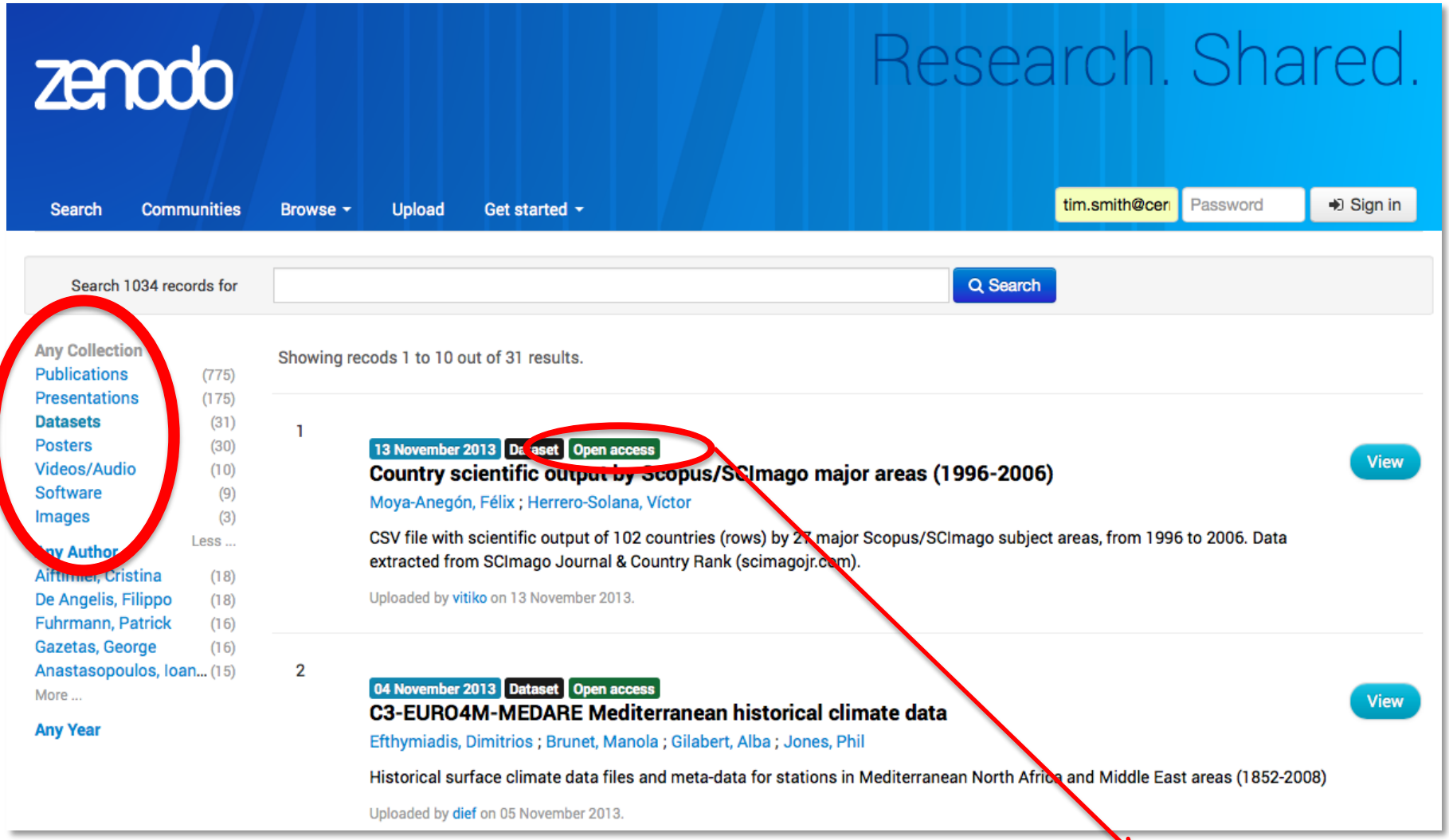
Export



Direct community upload

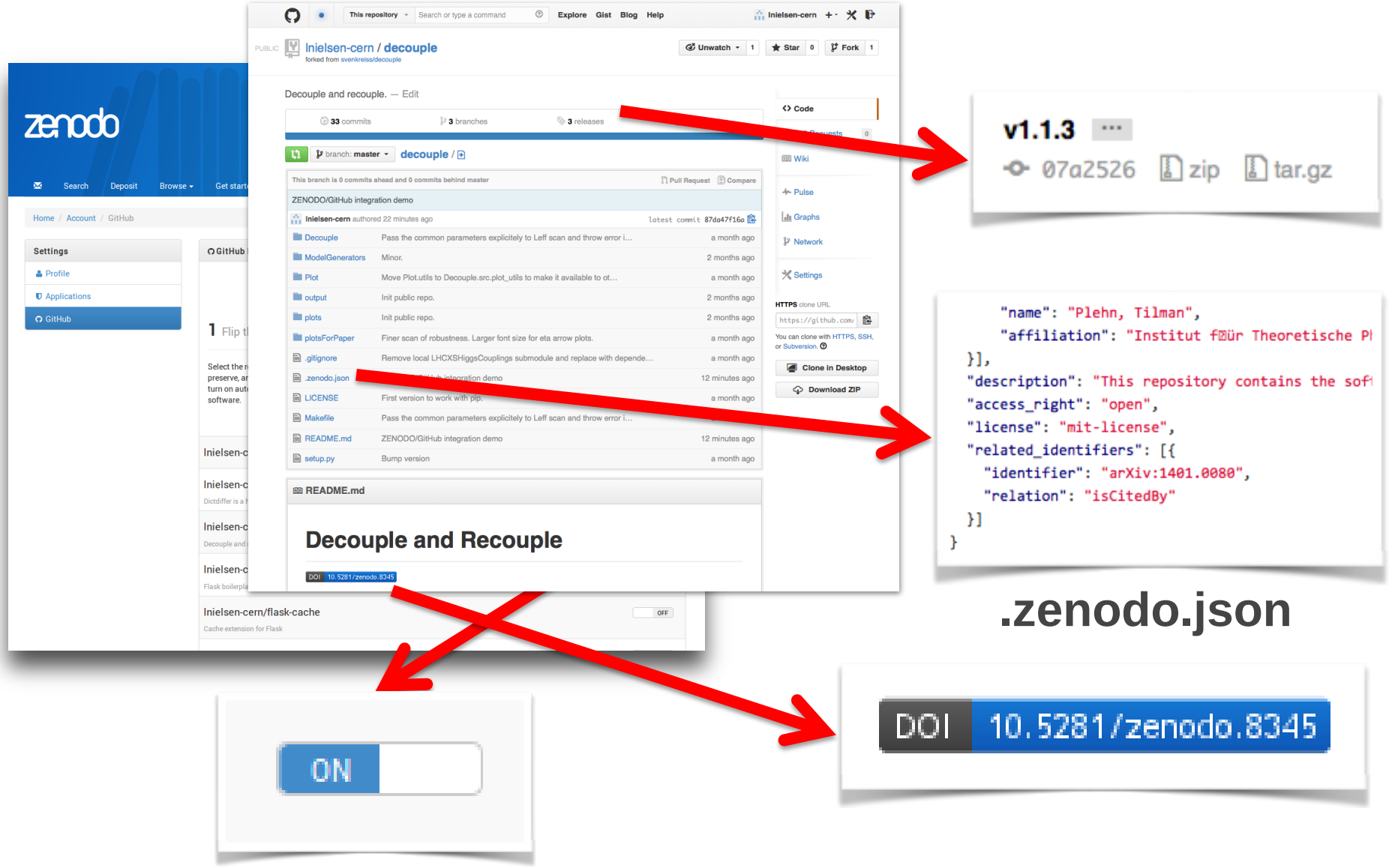


Research Repository



Closed access | Embargoed access

Zenodo – GitHub bridge



Code ↔ Data ↔ Paper

zenodo

Research. Shared.

Search Communities Browse Uplo

07 March 2014

decouple software ass
arXiv:1401.0080

Cranmer, Kyle ; Kreiss, Sven

(show affiliations)

This repository contains the software implementation
Higgs Coupling Measurements (Cranmer, Kreiss,
It contains tools to apply the discussed methods to ne
recreate the plots in the paper.

A demo for the recoupling stage where the effective lik
readily provided is at decoupledDemo.

Files

Name	Date
decouple-v1.2.5.zip	08 Mar 2014

INSPIRE HEP

Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPN

Information Citations (0)

Description:

This repository contains
Measurements (Cranm
to new models and cont

A demo for the recouplir
at decoupledDemo.

This dataset complen
A Novel Approach to

Record created 2014-03-12, la

Link to Zen
Link to Git

INSPIRE HEP

Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

Information References (166) Citations (0) Files Plots HepData

A Novel Approach to Higgs Coupling Measurements

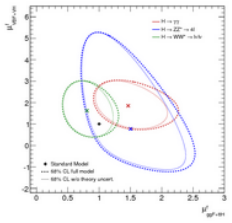
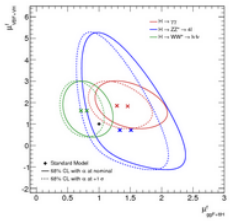
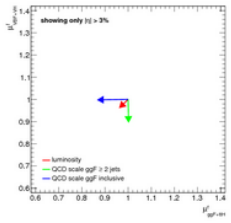
Kyle Cranmer, Sven Kreiss (New York U., CCPP) , David Lopez-Val (Louvain U., CP3) , Tilman Plehn (U. Heidelberg, ITP)

Dec 30, 2013 - 39 pages

e-Print: arXiv:1401.0080 [hep-ph] | PDF

Abstract (arXiv)
We develop a technique to present Higgs coupling measurements, which decouple the poorly defined theoretical uncertainties associated to inclusive and exclusive cross section predictions. The technique simplifies the combination of multiple measurements and can be used in a more general setting. We illustrate the approach with toy LHC Higgs coupling measurements and a collection of new physics models.

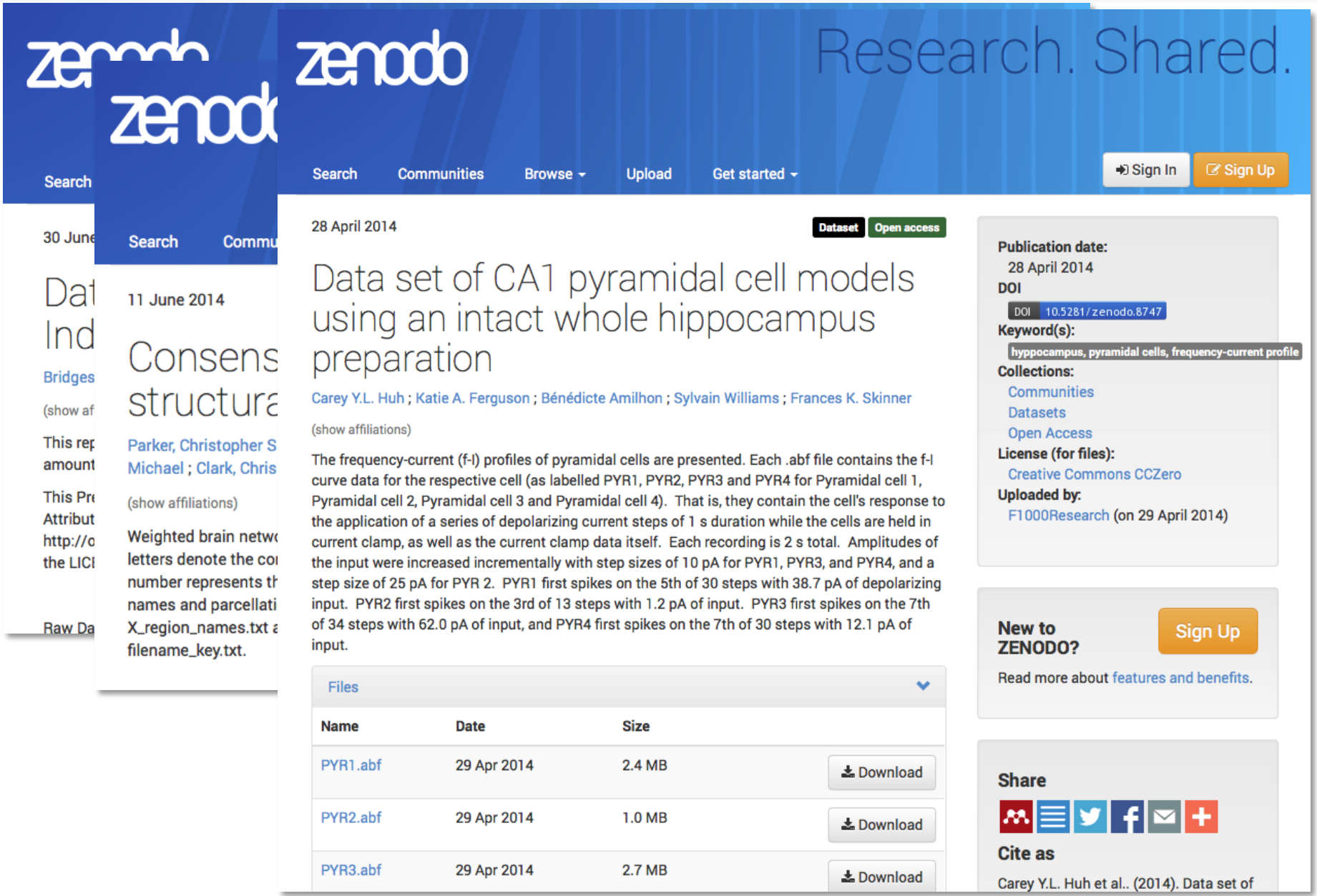
Note: 39 pages, 12 figures
Keyword(s): INSPIRE: * Automatic Keywords * | coupling: Higgs | CERN LHC Coll | new physics | decoupling

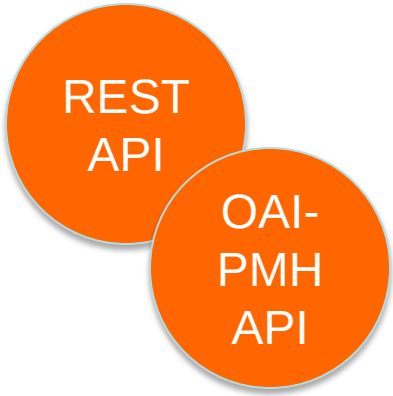
Show more plots

Record created 2014-01-03, last modified 2014-02-23

Example Zenodo Records



Integration Ideas



- Nordic Countries: clinical trial data
- Switzerland: life-science laboratory experiment notebook backend
- US: Virtual machines to run stored code on data

Conclusions

- *Information is a valuable asset that is multiplied when it is shared*
- Zenodo offered to help make Open Data a reality
- Preserving forever is an interesting challenge
  - Involving technological and sustainability innovation
  - The LHC started in 1984 … Higgs Boson discovery 2012
  - Let's explore the challenge together!

Tim.Smith@cern.ch
http://www.cern.ch

zenodo

Research. Shared.

http://zenodo.org
@zenodo_org

CC BY