CERN and HEPData

Sünje Dallmeier-Tiessen, Salvatore MELE, CERN

November 24th, 2017

Who are we

Salvatore Mele

Data analysis at LEP. In charge of Open Access at CERN (SCOAP3, INSPIRE et al).

Helped (re-)build current INSPIRE collaboration, service, team and partnerships with physicists and developers.

Governance positions on scientific/policy information initiatives (ORCID, DataCite,…)

Sünje Dallmeier-Tiessen

Information scientist. In charge of data curation at CERN

CERN Open Data, CERN Analysis Preservation, REANA, with teams of developers and information scientists.

Many advisory positions on data journals, repositories, initiatives globally.

INSPIRE Collaboration

CERN

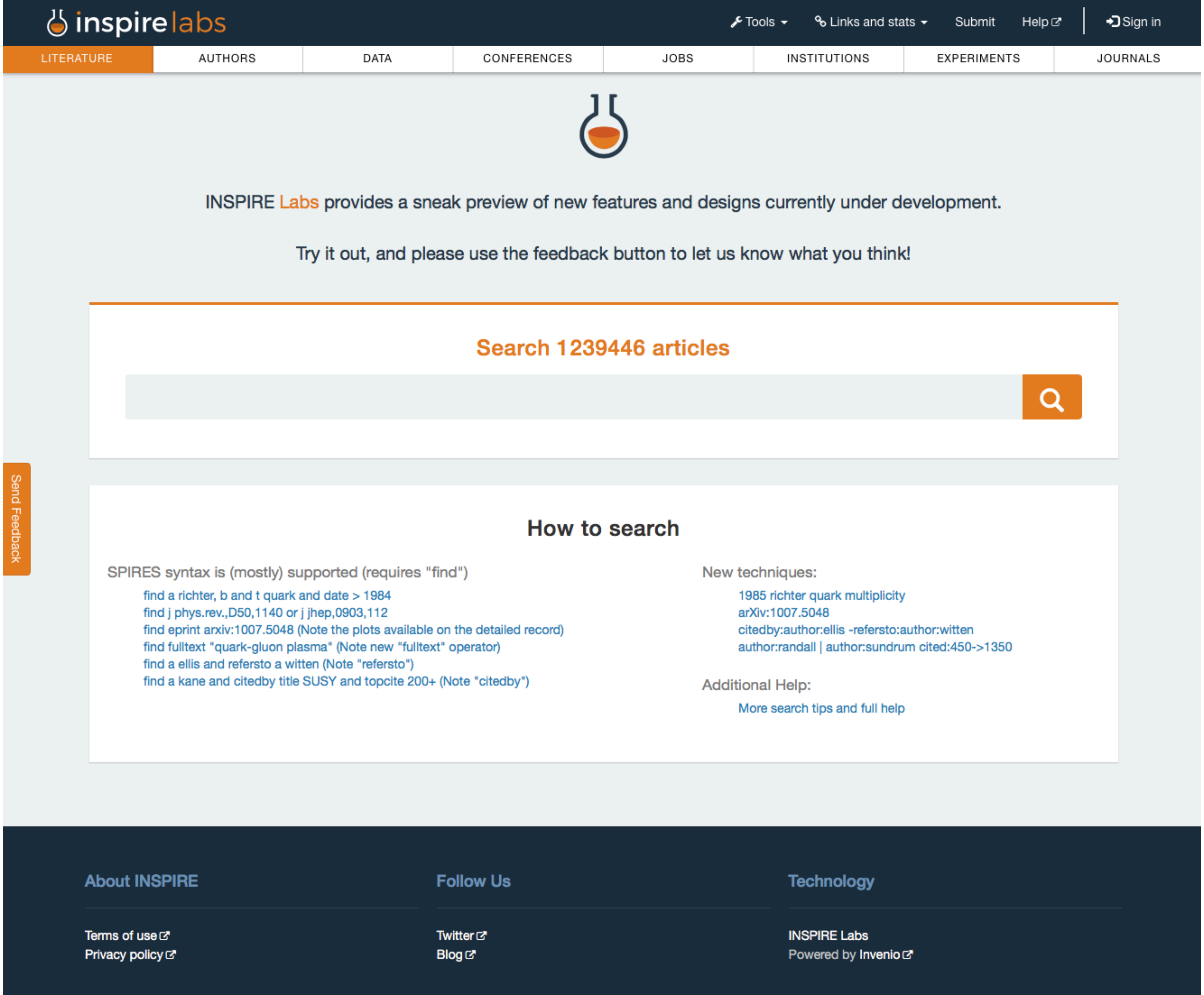
DESY

Fermilab

Institute of High Energy Physics
Chinese Academy of Sciences

SLAC
NATIONAL ACCELERATOR LABORATORY

- CERN hosts the INSPIRE development team (6 engineers and one product manager) and database, service and machines
- All partners contribute in different parts and amounts to the workflow of ingestion and curation of content
  - getting citations right,
  - getting papers into the right author profiles,
  - disambiguating by hand some of those,
  - checking obscure journals for relevant content,
  - Jobs, conferences,….
- (Initially) separate bi-lateral agreement with France and Japan for specific national curation projects



INSPIRE Labs provides a sneak preview of new features and designs currently under development.

Try it out, and please use the feedback button to let us know what you think!

Search 1239446 articles

How to search

SPIRES syntax is (mostly) supported (requires "find")

find a richter, b and t quark and date > 1984
find j phys.rev.,D50,1140 or j jhep,0903,112
find eprint arxiv:1007.5048 (Note the plots available on the detailed record)
find fulltext "quark-gluon plasma" (Note new "fulltext" operator)
find a ellis and refersto a witten (Note "refersto")
find a kane and citedby title SUSY and topcite 200+ (Note "citedby")

New techniques:

1985 richter quark multiplicity
arXiv:1007.5048
citedby:author:ellis -refersto:author:witten
author:randall | author:sundrum cited:450->1350

Additional Help:

More search tips and full help

About INSPIRE

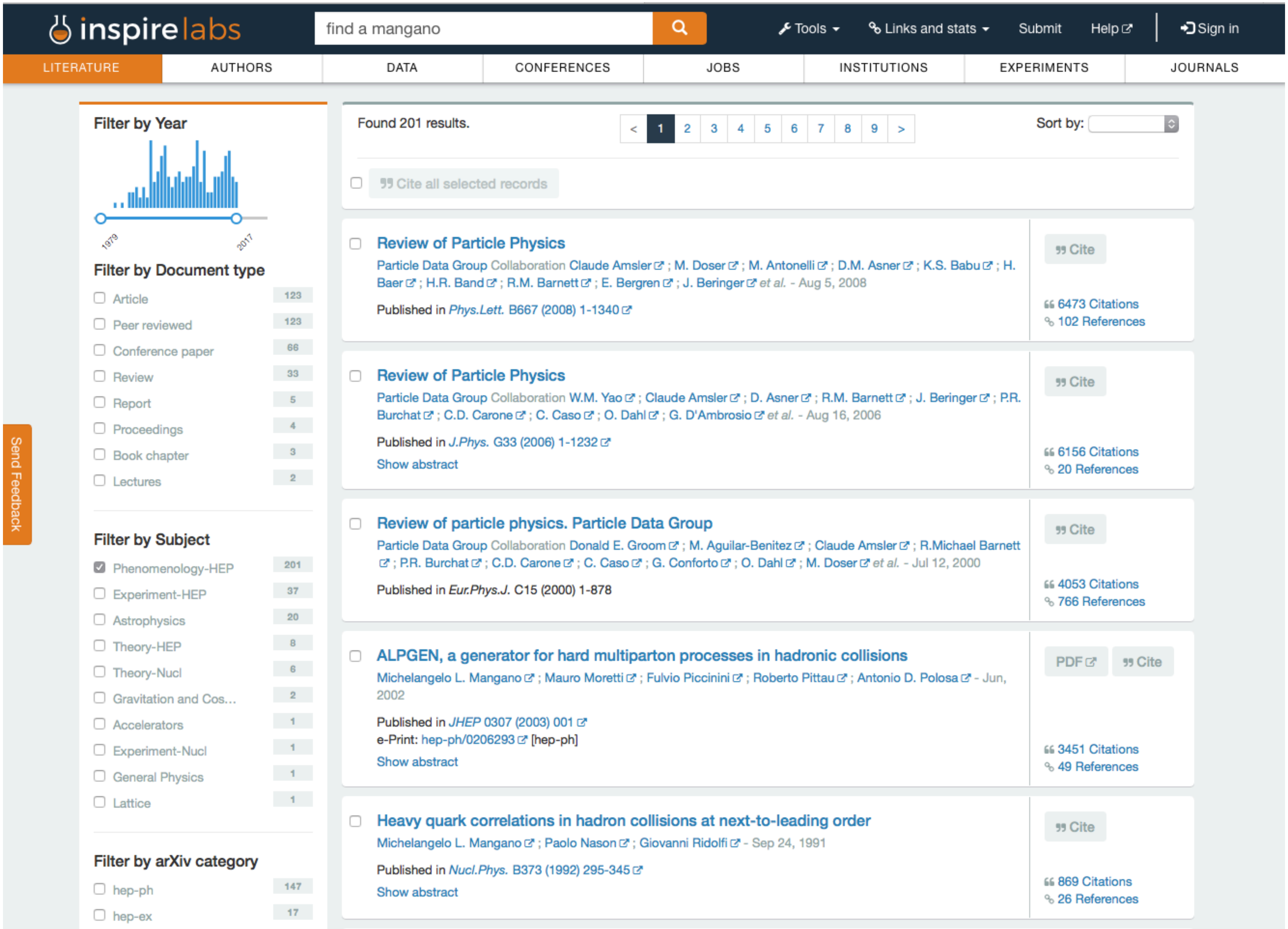
Terms of use
Privacy policy

Follow Us

Twitter
Blog

Technology

INSPIRE Labs
Powered by Invenio



find a mangano

Found 201 results.

Filter by Year

Filter by Document type

□ Article	123
□ Peer reviewed	123
□ Conference paper	66
□ Review	33
□ Report	5
□ Proceedings	4
□ Book chapter	3
□ Lectures	2

Filter by Subject

☑ Phenomenology-HEP	201
□ Experiment-HEP	37
□ Astrophysics	20
□ Theory-HEP	8
□ Theory-Nucl	6
□ Gravitation and Cos...	2
□ Accelerators	1
□ Experiment-Nucl	1
□ General Physics	1
□ Lattice	1

Filter by arXiv category

□ hep-ph	147
□ hep-ex	17

□ **Review of Particle Physics**
Particle Data Group Collaboration Claude Amsler ; M. Doser ; M. Antonelli ; D.M. Asner ; K.S. Babu ; H. Baer ; H.R. Band ; R.M. Barnett ; E. Bergren ; J. Beringer *et al.* - Aug 5, 2008
Published in *Phys.Lett.* B667 (2008) 1-1340
6473 Citations
102 References

□ **Review of Particle Physics**
Particle Data Group Collaboration W.M. Yao ; Claude Amsler ; D. Asner ; R.M. Barnett ; J. Beringer ; P.R. Burchat ; C.D. Carone ; C. Caso ; O. Dahl ; G. D'Ambrosio *et al.* - Aug 16, 2006
Published in *J.Phys.* G33 (2006) 1-1232
Show abstract
6156 Citations
20 References

□ **Review of particle physics. Particle Data Group**
Particle Data Group Collaboration Donald E. Groom ; M. Aguilar-Benitez ; Claude Amsler ; R.Michael Barnett ; P.R. Burchat ; C.D. Carone ; C. Caso ; G. Conforto ; O. Dahl ; M. Doser *et al.* - Jul 12, 2000
Published in *Eur.Phys.J.* C15 (2000) 1-878
4053 Citations
766 References

□ **ALPGEN, a generator for hard multiparton processes in hadronic collisions**
Michelangelo L. Mangano ; Mauro Moretti ; Fulvio Piccinini ; Roberto Pittau ; Antonio D. Polosa - Jun, 2002
Published in *JHEP* 0307 (2003) 001
e-Print: hep-ph/0206293 [hep-ph]
Show abstract
3451 Citations
49 References

□ **Heavy quark correlations in hadron collisions at next-to-leading order**
Michelangelo L. Mangano ; Paolo Nason ; Giovanni Ridolfi - Sep 24, 1991
Published in *Nucl.Phys.* B373 (1992) 295-345
Show abstract
869 Citations
26 References

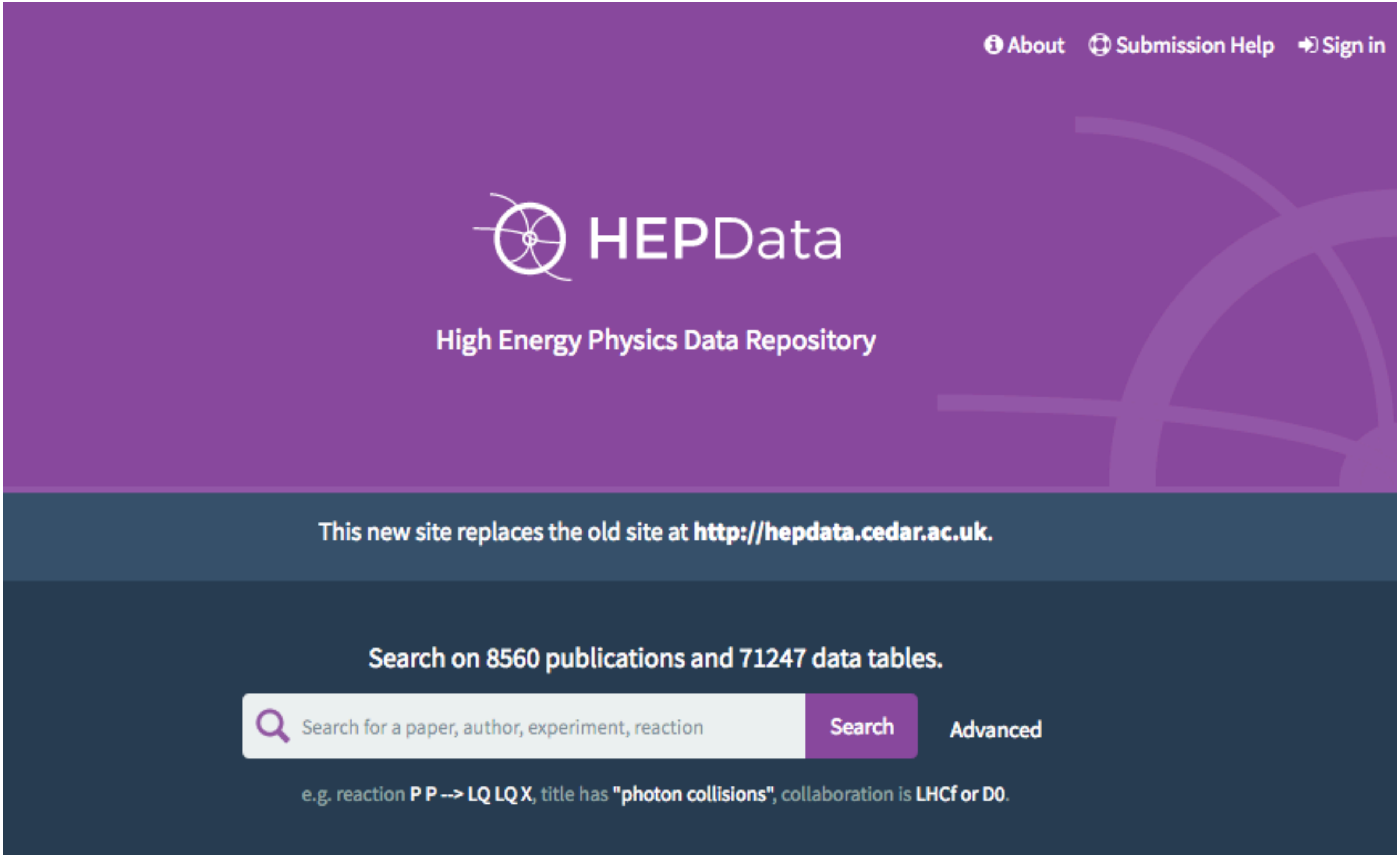
200K lines of code, machine learning author disambiguation, text mining to reduce hand curation, options for authors to improve information, beta soon, mostly done by 2019

SPIRES, INSPIRE, HEPData, CERN… a long story

- Long collaboration with HEPData – from SPIRES to INSPIRE: facilitating discovery/navigation through article-data links
- Hard to support deeper navigation from INSPIRE and the old HEPData infrastructure
- Hard to 'attribute', 'count' and 'cite' data (as the requests picked up)

"Rough" timeline

- 2014-2015: discussion on mutual benefits of Durham, CERN, INSPIRE, HEP community in modernising HEPData
- 2015-2016: CERN investment of a developer to create the 'new' HEPData customising CERN digital library technology (the same which is becoming a 'new' INSPIRE)
- April 2016 Workshop to present new HEPData at CERN: https://indico.cern.ch/event/512652/
- 2016+: CERN runs HEPData web service on as-is basis on its cloud infrastructure



High Energy Physics Data Repository

This new site replaces the old site at **http://hepdata.cedar.ac.uk**.

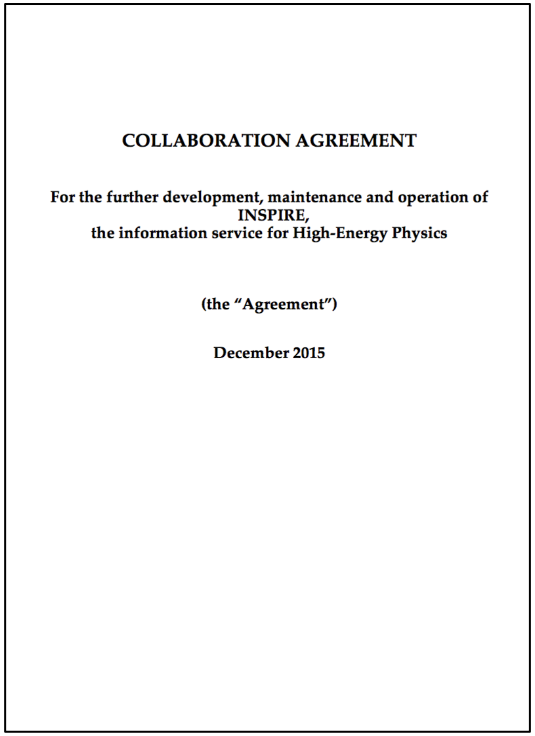
Search on 8560 publications and 71247 data tables.

e.g. reaction **P P --> LQ LQ X**, title has **"photon collisions"**, collaboration is **LHCf or D0**.

More formally: agreements and commitments

- Discussing an extension of the INSPIRE Collaboration (currently CERN, DESY, IHEP Beijing, Fermilab, SLAC) to formally include IPPP Durham, and clarify commitments to curate, develop, operate HEPData

- INSPIRE Collaboration has its MoU ('Collaboration Agreement') where each party commit to particular efforts, services. Aggregating data, and make it findable, is part of the core INSPIRE mission. This create the conditions for for IPPP accession.



COLLABORATION AGREEMENT

For the further development, maintenance and operation of INSPIRE,
the information service for High-Energy Physics

(the "Agreement")

December 2015

Relatively clear part:

- IPPP to commit to ingestion and curation of HEPData content
- CERN to commit to running the software as-is, and hosting the web service

Critical issues for long-term stability:

- No resources for software maintenance and necessary/desirable developments

Opportunities for a disciplinary data-repository today

Recent data repository evolution

- Open Science is strongly asserted by many funding agencies, (FAIR) access to data a main pillar. Many start to 'count' data among outputs to report. UK leading the way in requirements

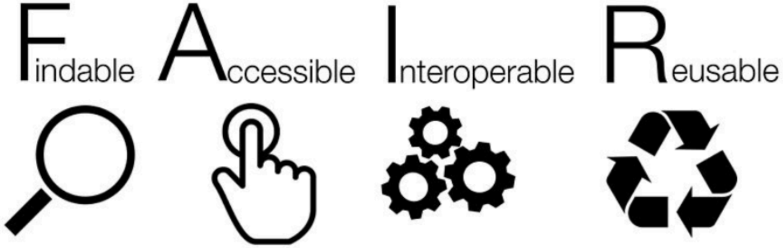
- Countless disciplinary and broad-band examples of data repositories, starting early 2000 and intensifying

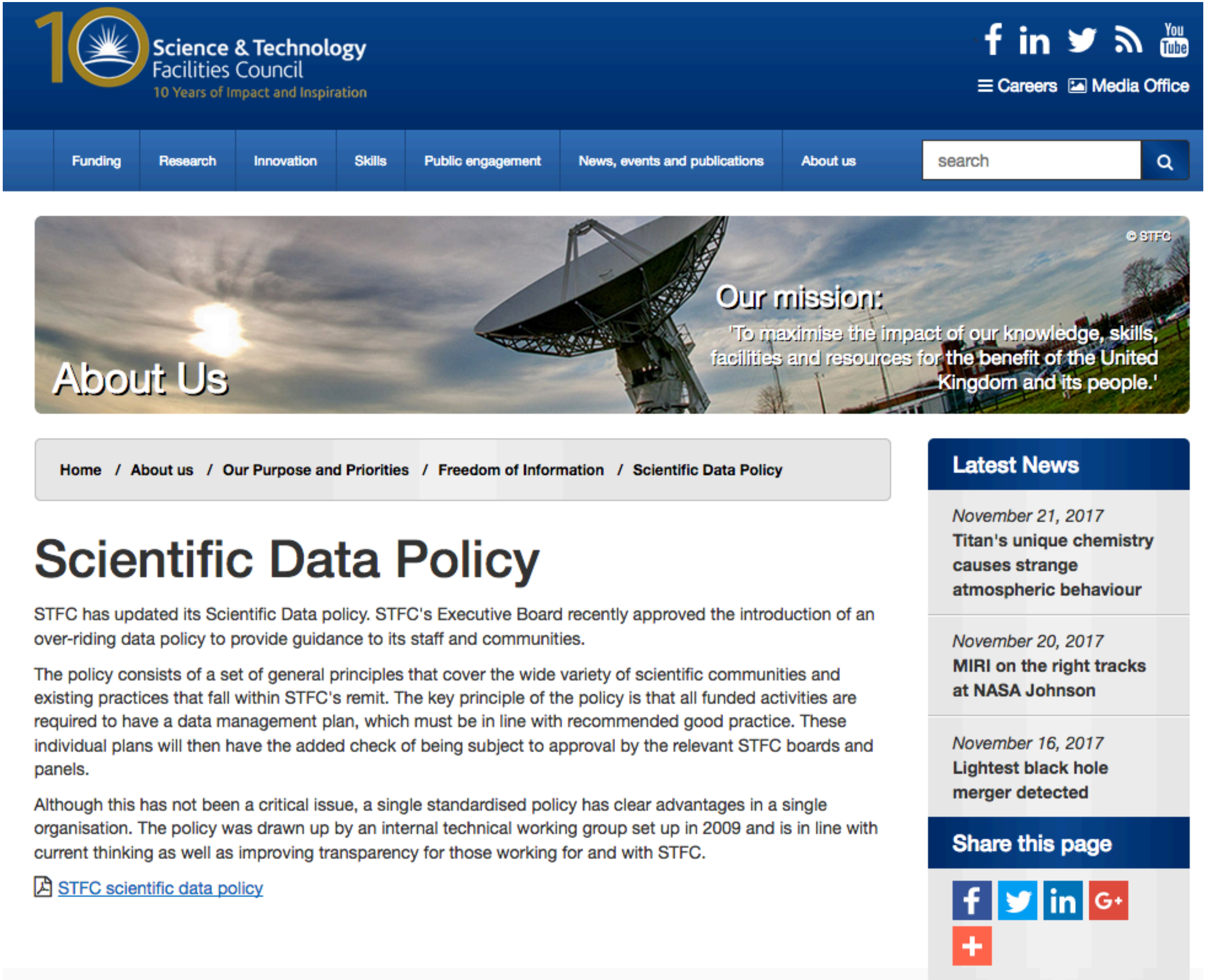
- Authoritative registry of research data repositories (re3data.org) counts 1930

Beyond the repository

- Publishers increasingly require 'data availability statements' and make data citation a standard

- More and more initiatives (Publishers, ORCID, funding agencies) expect article-data linking as standard




Science & Technology Facilities Council
10 Years of Impact and Inspiration
Careers
Media Office
Funding
Research
Innovation
Skills
Public engagement
News, events and publications
About us
search
© STFC
About Us

Our mission:
'To maximise the impact of our knowledge, skills, facilities and resources for the benefit of the United Kingdom and its people.'

Home / About us / Our Purpose and Priorities / Freedom of Information / Scientific Data Policy

Scientific Data Policy

STFC has updated its Scientific Data policy. STFC's Executive Board recently approved the introduction of an over-riding data policy to provide guidance to its staff and communities.

The policy consists of a set of general principles that cover the wide variety of scientific communities and existing practices that fall within STFC's remit. The key principle of the policy is that all funded activities are required to have a data management plan, which must be in line with recommended good practice. These individual plans will then have the added check of being subject to approval by the relevant STFC boards and panels.

Although this has not been a critical issue, a single standardised policy has clear advantages in a single organisation. The policy was drawn up by an internal technical working group set up in 2009 and is in line with current thinking as well as improving transparency for those working for and with STFC.

STFC scientific data policy

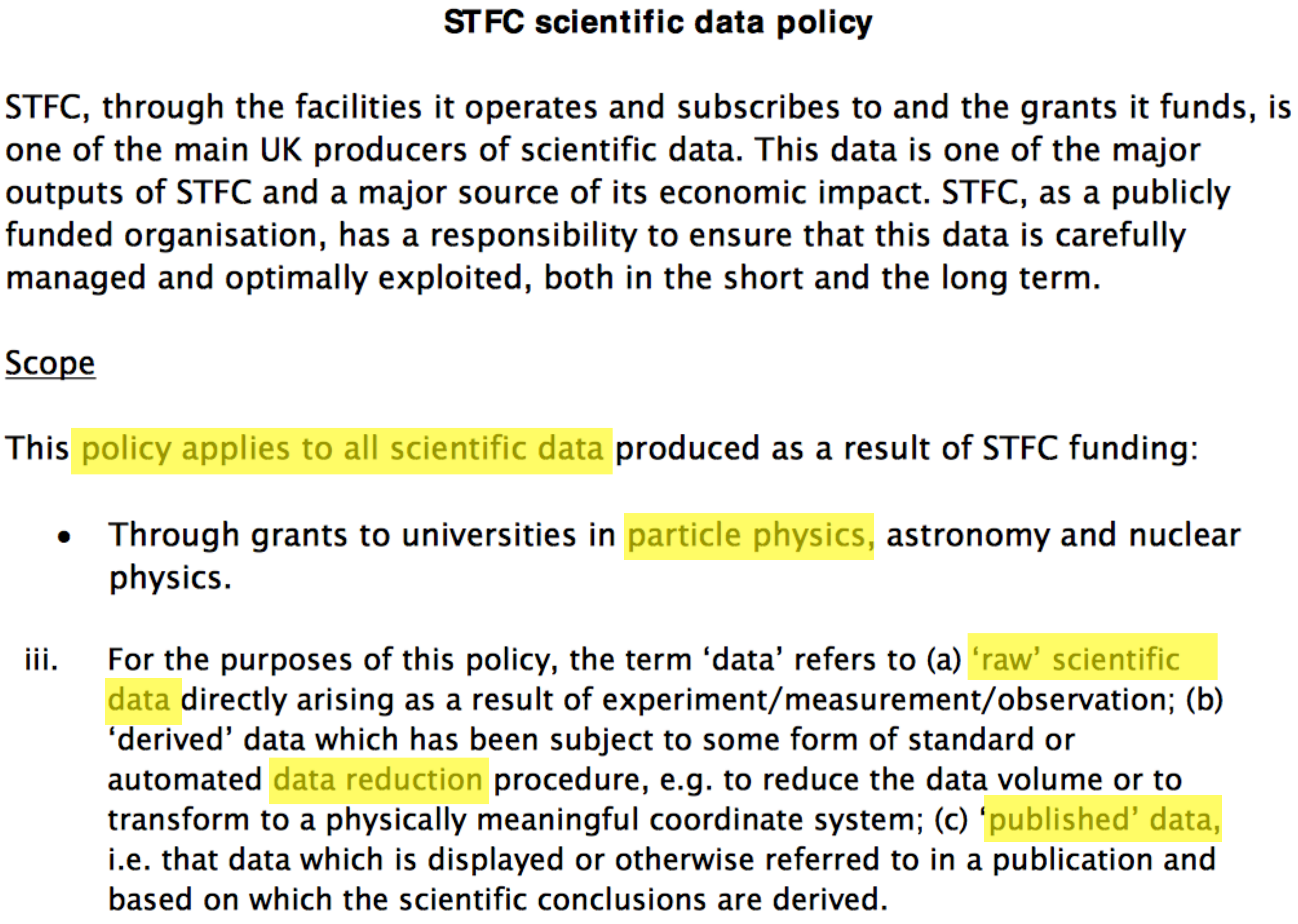
Latest News

November 21, 2017
Titan's unique chemistry causes strange atmospheric behaviour

November 20, 2017
MIRI on the right tracks at NASA Johnson

November 16, 2017
Lightest black hole merger detected

Beyond tables and figures !



STFC scientific data policy

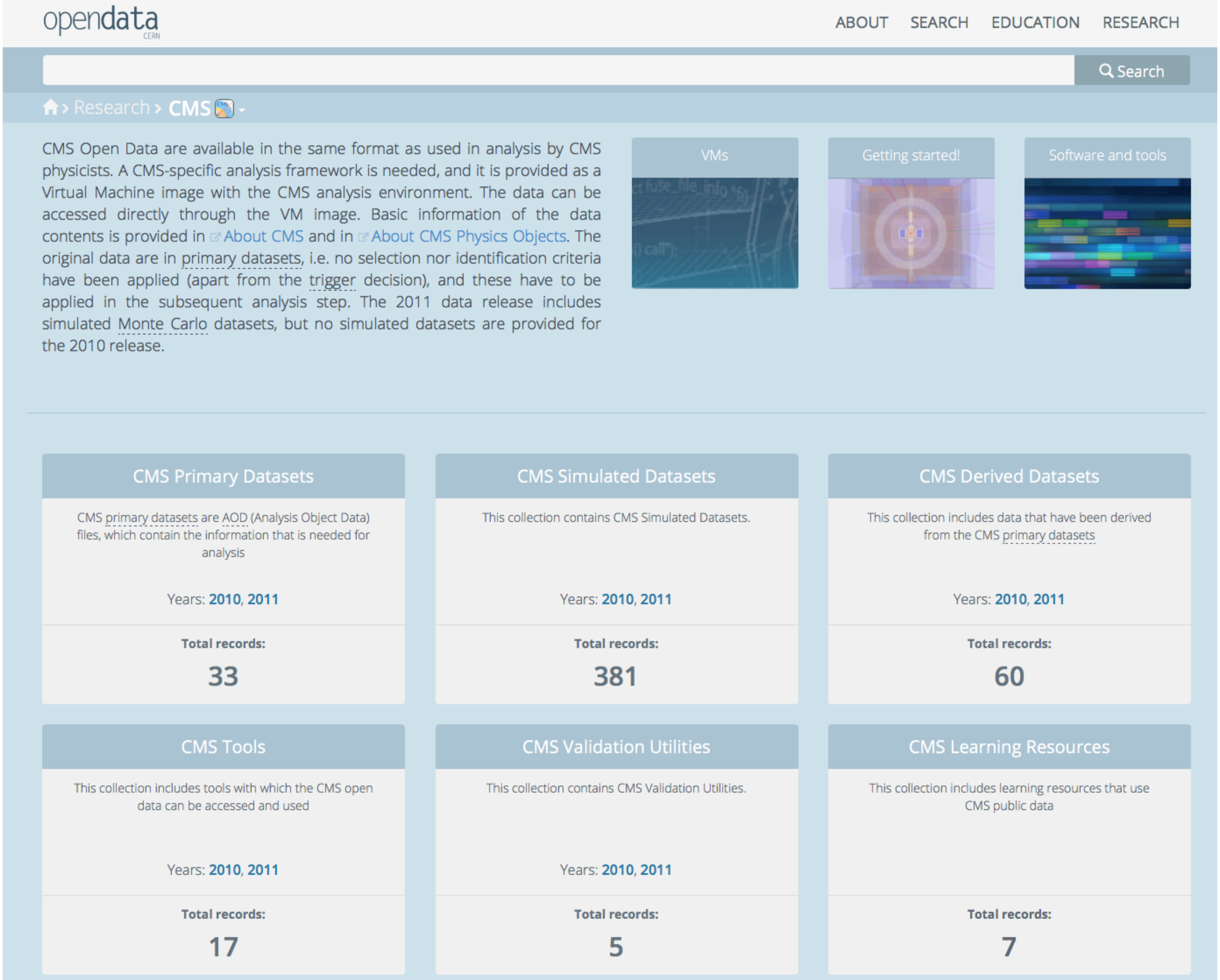
STFC, through the facilities it operates and subscribes to and the grants it funds, is one of the main UK producers of scientific data. This data is one of the major outputs of STFC and a major source of its economic impact. STFC, as a publicly funded organisation, has a responsibility to ensure that this data is carefully managed and optimally exploited, both in the short and the long term.

Scope

This policy applies to all scientific data produced as a result of STFC funding:

- Through grants to universities in particle physics, astronomy and nuclear physics.

iii. For the purposes of this policy, the term ‘data’ refers to (a) ‘raw’ scientific data directly arising as a result of experiment/measurement/observation; (b) ‘derived’ data which has been subject to some form of standard or automated data reduction procedure, e.g. to reduce the data volume or to transform to a physically meaningful coordinate system; (c) ‘published’ data, i.e. that data which is displayed or otherwise referred to in a publication and based on which the scientific conclusions are derived.



CMS Open Data are available in the same format as used in analysis by CMS physicists. A CMS-specific analysis framework is needed, and it is provided as a Virtual Machine image with the CMS analysis environment. The data can be accessed directly through the VM image. Basic information of the data contents is provided in About CMS and in About CMS Physics Objects. The original data are in primary datasets, i.e. no selection nor identification criteria have been applied (apart from the trigger decision), and these have to be applied in the subsequent analysis step. The 2011 data release includes simulated Monte Carlo datasets, but no simulated datasets are provided for the 2010 release.

CMS Primary Datasets

CMS primary datasets are AOD (Analysis Object Data) files, which contain the information that is needed for analysis

Years: 2010, 2011

Total records:

33

CMS Simulated Datasets

This collection contains CMS Simulated Datasets.

Years: 2010, 2011

Total records:

381

CMS Derived Datasets

This collection includes data that have been derived from the CMS primary datasets

Years: 2010, 2011

Total records:

60

CMS Tools

This collection includes tools with which the CMS open data can be accessed and used

Years: 2010, 2011

Total records:

17

CMS Validation Utilities

This collection contains CMS Validation Utilities.

Years: 2010, 2011

Total records:

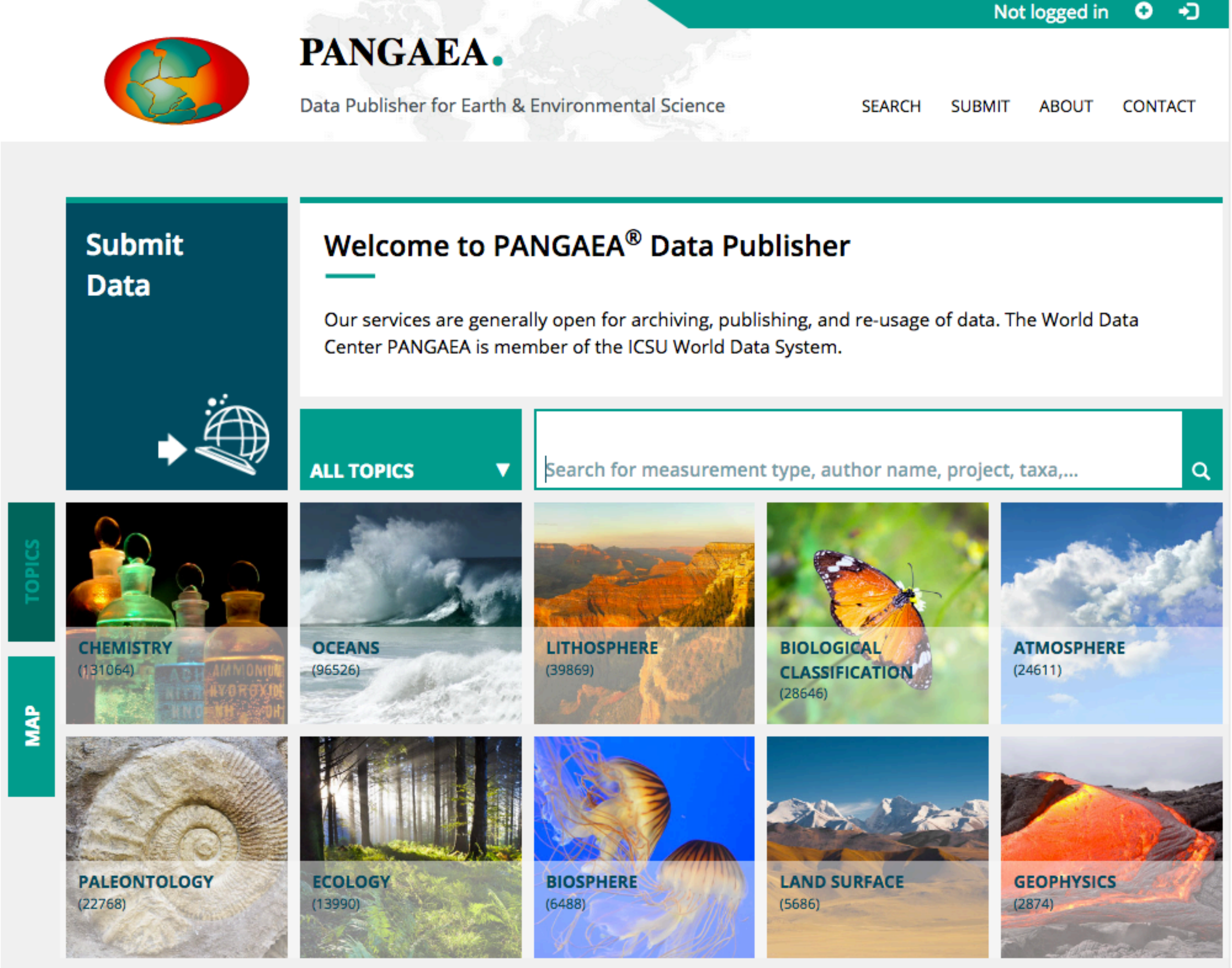
5

CMS Learning Resources

This collection includes learning resources that use CMS public data

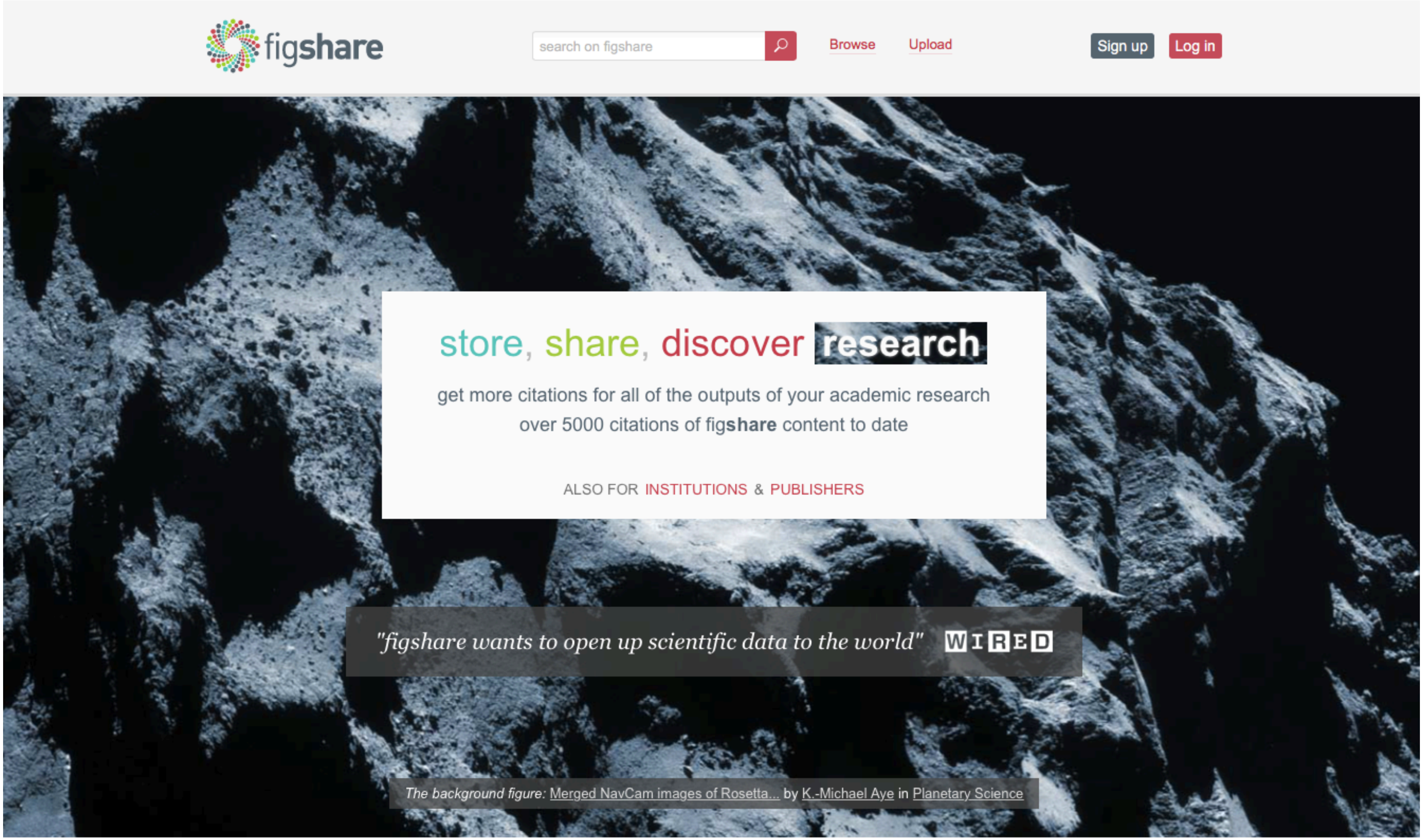
Total records:

7



Welcome to PANGAEA® Data Publisher

Our services are generally open for archiving, publishing, and re-usage of data. The World Data Center PANGAEA is member of the ICSU World Data System.



store, share, discover research

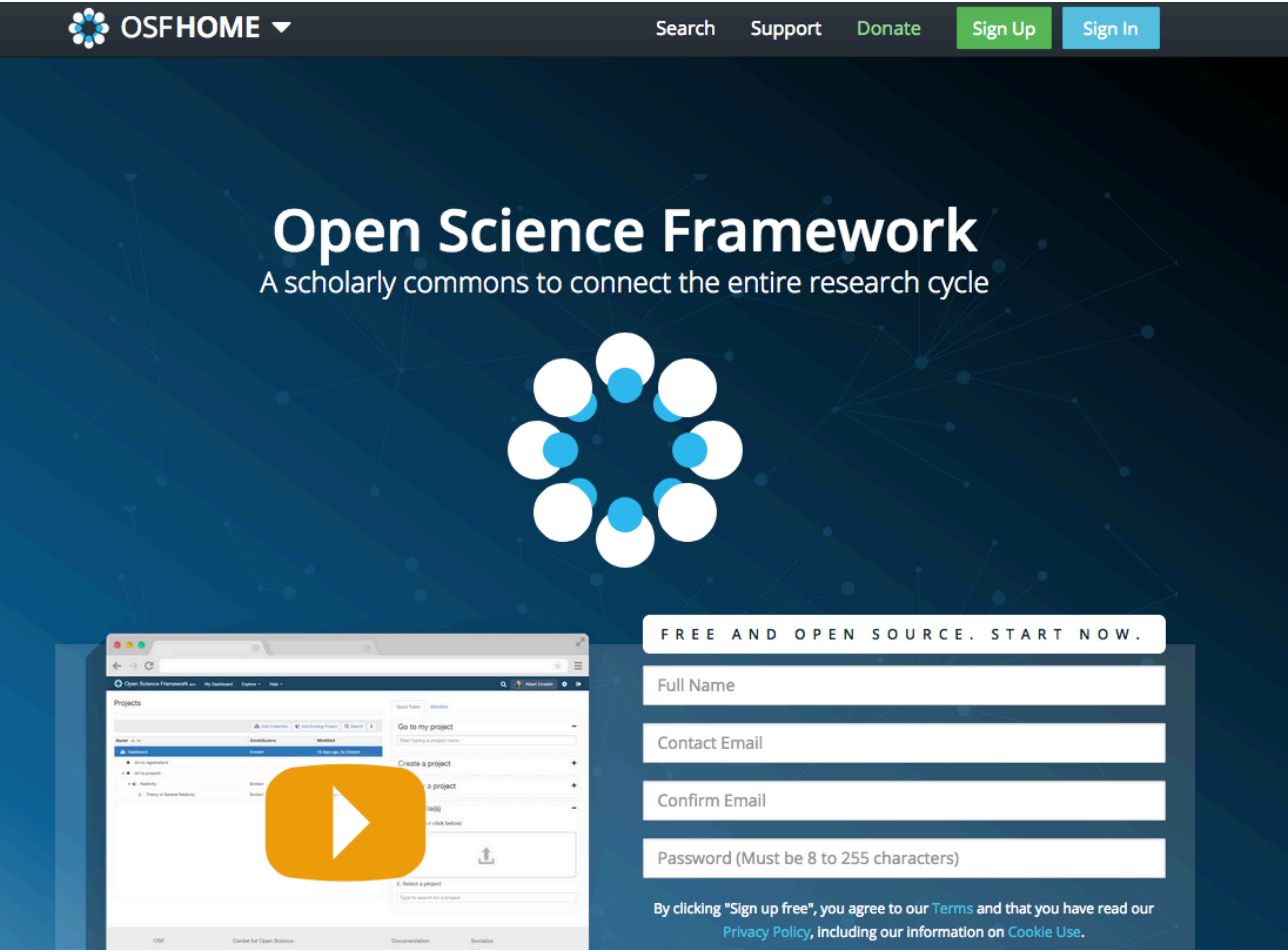
get more citations for all of the outputs of your academic research
over 5000 citations of figshare content to date

ALSO FOR INSTITUTIONS & PUBLISHERS

"figshare wants to open up scientific data to the world"

The background figure: Merged NavCam images of Rosetta... by K.-Michael Aye in Planetary Science

simplifying your research workflow



Open Science Framework

A scholarly commons to connect the entire research cycle

FREE AND OPEN SOURCE. START NOW.

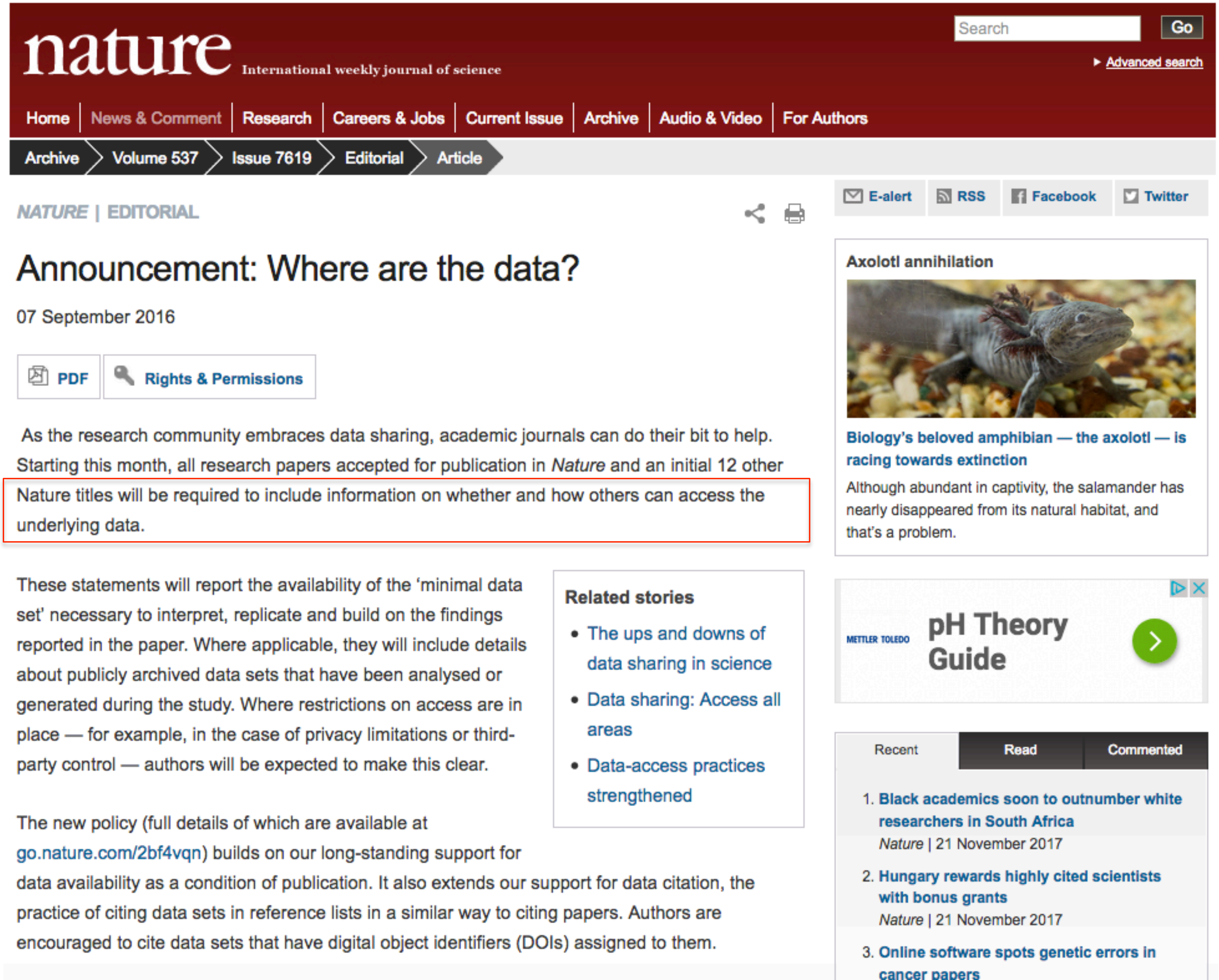
Full Name

Contact Email

Confirm Email

Password (Must be 8 to 255 characters)

By clicking "Sign up free", you agree to our Terms and that you have read our Privacy Policy, including our information on Cookie Use.


nature
International weekly journal of science
Search
Go
Advanced search
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors
Archive > Volume 537 > Issue 7619 > Editorial > Article
E-alert
RSS
Facebook
Twitter
PDF
Rights & Permissions

NATURE | EDITORIAL

Announcement: Where are the data?

07 September 2016

As the research community embraces data sharing, academic journals can do their bit to help. Starting this month, all research papers accepted for publication in *Nature* and an initial 12 other Nature titles will be required to include information on whether and how others can access the underlying data.

These statements will report the availability of the 'minimal data set' necessary to interpret, replicate and build on the findings reported in the paper. Where applicable, they will include details about publicly archived data sets that have been analysed or generated during the study. Where restrictions on access are in place — for example, in the case of privacy limitations or third-party control — authors will be expected to make this clear.

Related stories

- The ups and downs of data sharing in science
- Data sharing: Access all areas
- Data-access practices strengthened

The new policy (full details of which are available at go.nature.com/2bf4vqn) builds on our long-standing support for data availability as a condition of publication. It also extends our support for data citation, the practice of citing data sets in reference lists in a similar way to citing papers. Authors are encouraged to cite data sets that have digital object identifiers (DOIs) assigned to them.

Axolotl annihilation

Biology's beloved amphibian — the axolotl — is racing towards extinction

Although abundant in captivity, the salamander has nearly disappeared from its natural habitat, and that's a problem.

1. **Black academics soon to outnumber white researchers in South Africa**
Nature | 21 November 2017
2. **Hungary rewards highly cited scientists with bonus grants**
Nature | 21 November 2017
3. **Online software spots genetic errors in cancer papers**

Leveraging HEPData for the future

The HEP Community (because of funding agencies, and for its own good!) needs a place to deposit (small size) data beyond figures and plots, growing slowly from the current HEPData service and philosophy.

- INSPIRE originally started offering some small-scale ‘data deposit solution’. Not scalable, not appropriate, and confusing with INSPIRE central role of aggregator of HEP many sources (e.g. submit preprints to arXiv, then publish on JHEP, then find all info and links on INSPIRE, then push info to ORCID)

- Broadband/institution solutions (Figshare, OSF/COS, university data repositories in libraries…) means information is scattered and not tailored to HEP needs, hard to find, aggregate, reuse… effectively not FAIR

- HEPData is uniquely positioned to leverage this opportunity:
  - current modern software stack
  - trusted ‘brand’ in the community and full community support
  - use the coming STFC funding round to propose this new role
  - serve UK experimental HEP community and beyond

Concrete advice for the bid (2FTE):

1. Development effort to adapt and grow software to broader content.

   Our advice/experience: 0.75 FTE developer - full-stack engineer with fluent python, Javascript, strong software architecture, interest in Open Science

2. Devise within the HEP experimental and phenomenology communities a policy for which content should be targeted/prioritised and build pilot services

   Our advice/experience: 0.50 FTE information scientist with education/work experience in curating/developing data repositories

3. Interface with all other system and actors in the field (INSPIRE, ORCID, DataCite, Publishers, HEP code/MC/tool initiative) to align and complement services

   Our advice/experience: 0.25 FTE information scientist with understanding of wider context, with punctual guidance from physicists in the institute/community

4. (Of course) continued curation effort for existing content types and maintenance of current software stack

   Our advice/experience: the remaining 0.25 FTE curator + 0.25 FTE developer