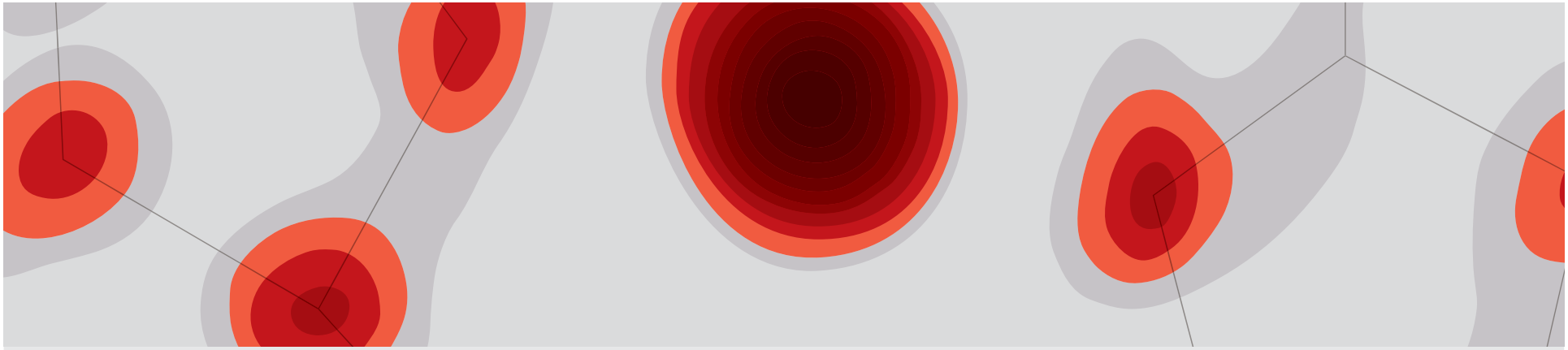
Part of SPRINGER NATURE



Beyond supplementary material:
sharing data effectively through repositories and data journals

Andrew L Hufton

Managing Editor, *Scientific Data*

andrew.hufton@nature.com

nature research

Science is about trust (with evidence)

1

Backing up your claims with evidence has always been a key part of science

Roger Bacon, 1214 – 1292

Opus Majus

“Theories supplied by reason should be verified by sensory data, aided by instruments, and corroborated by trustworthy witnesses

Opus Tertium

“The strongest argument proves nothing so long as the conclusions are not verified by experience.

Without proper data handling things can go wrong…

RETRACTION

doi:10.1038/nature12897

Retraction: The NAD-dependent deacetylase SIRT2 is required for programmed necrosis

Nisha Narayan, In Hye Lee, Ronen Borenstein, Junhui Sun, Renee Wong, Guang Tong, Maria M. Fergusson, Jie Liu, Ilsa I. Rovira, Hwei-Ling Cheng, Guanghui Wang, Marjan Gucek, David Lombard, Fredrick W. Alt, Michael N. Sack, Elizabeth Murphy, Liu Cao & Toren Finkel

Nature **492,** 199–204 (2012); doi:10.1038/nature11700

We retract this Article because some of the data, specifically the data reported in Fig. 2 demonstrating an *in vitro* requirement for Sirt2 in TNF-α-mediated necroptosis, appears to be irreproducible. We and others have confirmed that Sirt2 and RIP3 interact, and we continue to believe that the absence of Sirt2 protects against ischaemic myocardial damage. Nonetheless, our inability to reproduce the data in Fig. 2 involving TNF-α-mediated necroptosis undermines our confidence in the scientific conclusions reporte
Although the matter is currently unde
the Article in its entirety, and regret a
have resulted from the paper's publi

Superconductivity in single crystals of the fullerene C_{70}

J. H. Schön, Ch. Kloc, T. Siegrist, M. Steigerwald, C. Svensson & B. Batlogg

Nature 413, 831–833 (2001).

This manuscript was, in part, the subject of an independent investigation[1] conducted at the behest of Bell Laboratories, Lucent Technologies. The independent committee reviewed concerns related to the validity of data associated with the device measurements described in the paper. As a result of the committee's findings, we are issuing a retraction of the paper. We note nevertheless that this paper may also contain some legitimate ideas and contributions. □

1. Beasley, M. R., Datta, S., Kogelnik, H., Kroemer, H. & Monroe, D. Report of the Investigation

oauthors.
merican

RETRACTION

DNA sequencing using electrical conductance measurements of a DNA polymerase

Yu-Shiun Chen, Chia-Hui Lee, Meng-Yen Hung, Hsu-An Pan, Jin-Chern Chiou and G. Steven Huang

Nature Nanotechnology 8, 452–458 (2013); published online 5 May 2013; corrected after print 11 July 2013 and 28 August 2013; retracted after print 3 June 2015.

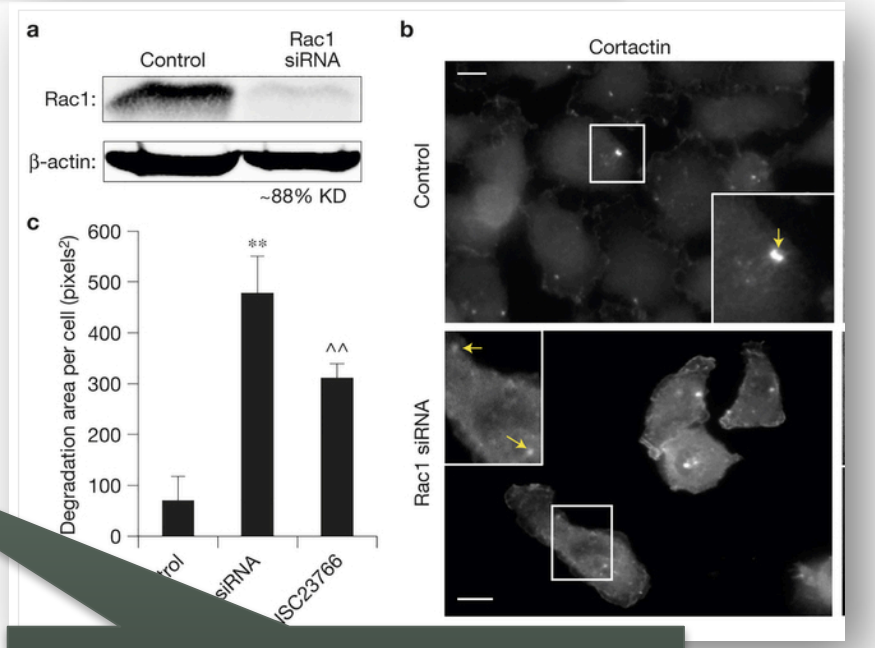
Significant concerns were raised about the validity of the data reported in this work shortly after publication. After an internal inquiry, a formal investigation was launched at the National Chiao Tung University, which focused on the reproducibility of the data. The results of the work could not be reproduced in a reasonable timeframe, and the authors could not provide the investigating committee with a complete set of raw data for the original experiments. The authors Y.-S. C., J.-C. C. and G. S. H. have therefore agreed to retract the manuscript; C.-H. L., M.-Y. H. and H.-A. P. did not respond to the journal's attempts to contact them about this retraction.

Raising reporting standards for data description

Checklist to improve figure legends and reporting

http://www.nature.com/authors/policies/checklist.pdf

(a) Western blot of cell lysates of control and Rac1-siRNA-treated MTLn3 cells, blotted for Rac1 and β-actin. **A representative image is shown from 3 blots**. **(b)** MTLn3 cells transfected with control or Rac1 siRNA and plated on Alexa-405-conjugated gelatin overnight. Arrows point to invadopodia and sites of degradation. Scale bars, 10 µm. Representative image sets are shown from 50 image sets each for the control and Rac1 siRNA. **(c)** Quantification of mean degradation area per cell from **b**, including Rac1 inhibitor NSC23766 treatment at 100 µM. ***n* = 60 fields for each condition, pooled from 5 independent experiments;** error bars are s.e.m. Student's *t*-test was used. ***P* = 0.00022,^ ^*P* = 0.01163 **Uncropped images** of blots are shown in Supplementary Fig. 9.



statement of replication

definition of *n*

definition of statistic tests

raw source data

Nature Cell Biology **16**, 571–583 (2014) doi:10.1038/ncb2972

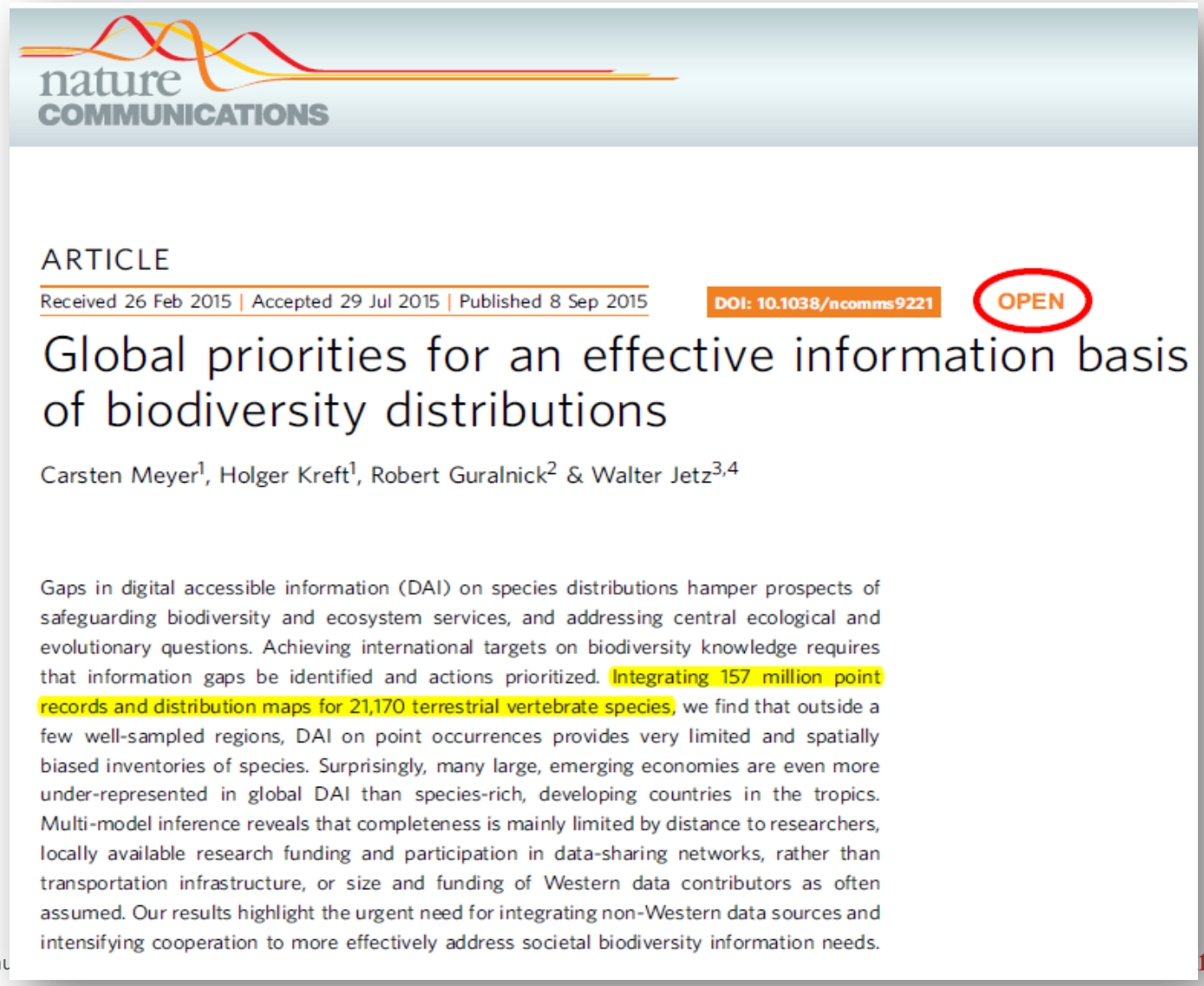
Large-scale data sharing is also not new

whole fields have been born from effective data sharing

- **Macroeconomics** grew from the sharing of government data with economists
- **Meteorology and climate science**: There has been international sharing of weather data for over 100 years
- **Bioinformatics:** Made possible by open sharing of protein structure and DNA sequence data

Sieber, J. E. **Data Sharing in Historical Perspective** (2015)
http://www.socialsciencespace.com/2015/09/data-sharing-in-historical-perspective/
Hufton, A. L. **Sharing the structures** doi:10.1038/nature13369 (2014)

Data sharing advances science



nature COMMUNICATIONS

ARTICLE

Received 26 Feb 2015 | Accepted 29 Jul 2015 | Published 8 Sep 2015

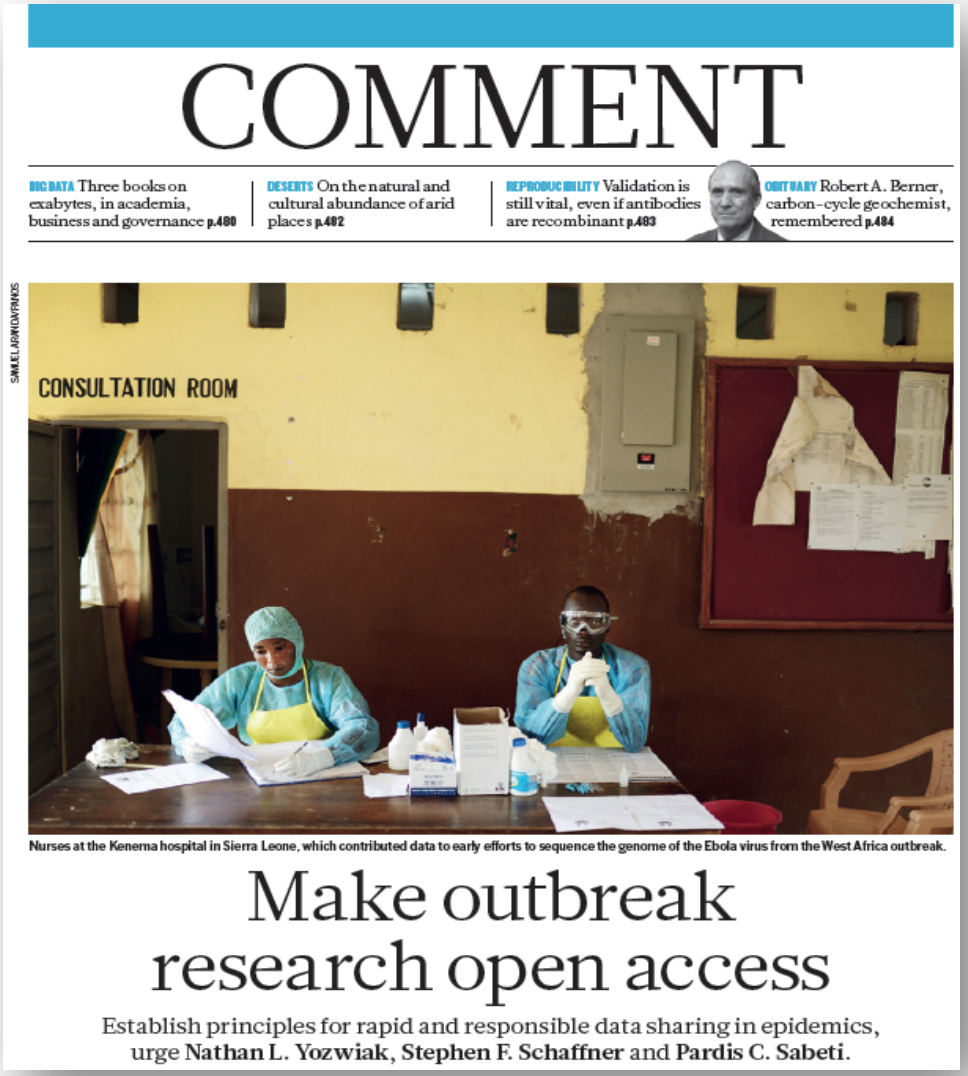
DOI: 10.1038/ncomms9221 OPEN

Global priorities for an effective information basis of biodiversity distributions

Carsten Meyer[1], Holger Kreft[1], Robert Guralnick[2] & Walter Jetz[3,4]

Gaps in digital accessible information (DAI) on species distributions hamper prospects of safeguarding biodiversity and ecosystem services, and addressing central ecological and evolutionary questions. Achieving international targets on biodiversity knowledge requires that information gaps be identified and actions prioritized. Integrating 157 million point records and distribution maps for 21,170 terrestrial vertebrate species, we find that outside a few well-sampled regions, DAI on point occurrences provides very limited and spatially biased inventories of species. Surprisingly, many large, emerging economies are even more under-represented in global DAI than species-rich, developing countries in the tropics. Multi-model inference reveals that completeness is mainly limited by distance to researchers, locally available research funding and participation in data-sharing networks, rather than transportation infrastructure, or size and funding of Western data contributors as often assumed. Our results highlight the urgent need for integrating non-Western data sources and intensifying cooperation to more effectively address societal biodiversity information needs.

Data sharing is in the public interest



COMMENT

BIG DATA Three books on exabytes, in academia, business and governance p.480 | DESERTS On the natural and cultural abundance of arid places p.482 | REPRODUCIBILITY Validation is still vital, even if antibodies are recombinant p.483 | OBITUARY Robert A. Berner, carbon-cycle geochemist, remembered p.484

SAMUEL ARYEH/PANOS

CONSULTATION ROOM

Nurses at the Kenema hospital in Sierra Leone, which contributed data to early efforts to sequence the genome of the Ebola virus from the West Africa outbreak.

Make outbreak research open access

Establish principles for rapid and responsible data sharing in epidemics, urge **Nathan L. Yozwiak, Stephen F. Schaffner** and **Pardis C. Sabeti.**

- We encourage our authors to share data related to public health emergencies as soon as possible, even *prior* to peer review or publication.

Nature **518,** 477–479 (2015)

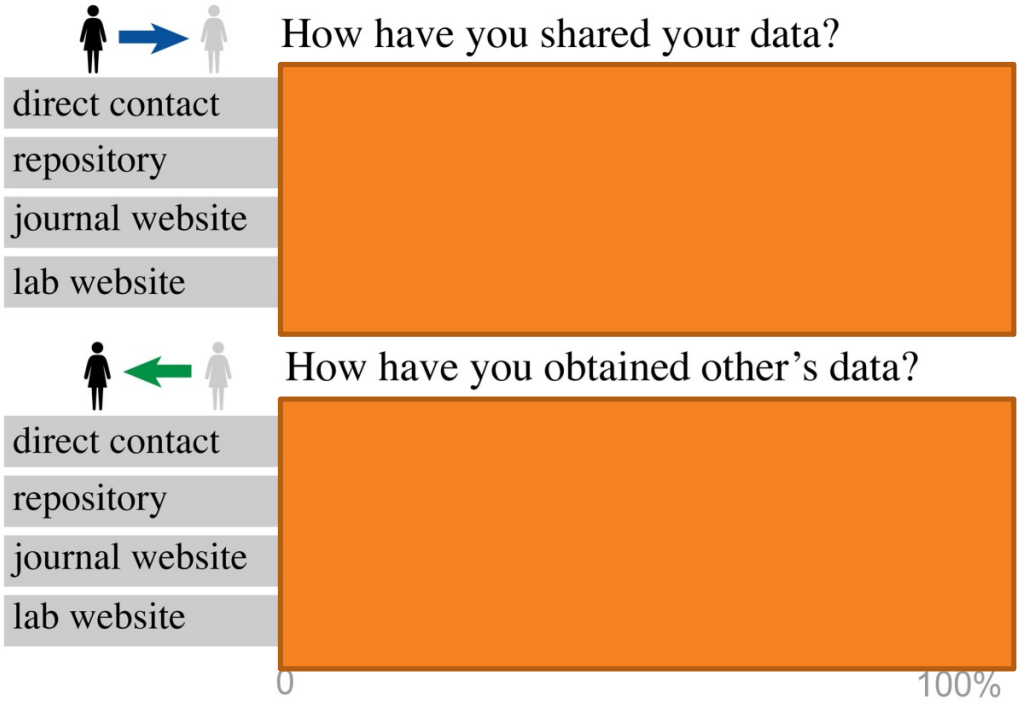
nature research

Data sharing – everyone's doing it

2

The current situation

- Most researchers are sharing data, and using the data of others
- Direct contact between researchers (on request) is a common way of sharing data
- Repositories are second most common method of sharing



Kratz JE, Strasser C (2015) Researcher Perspectives on Publication and Peer Review of Data. *PLoS ONE* 10(2): e0117619.

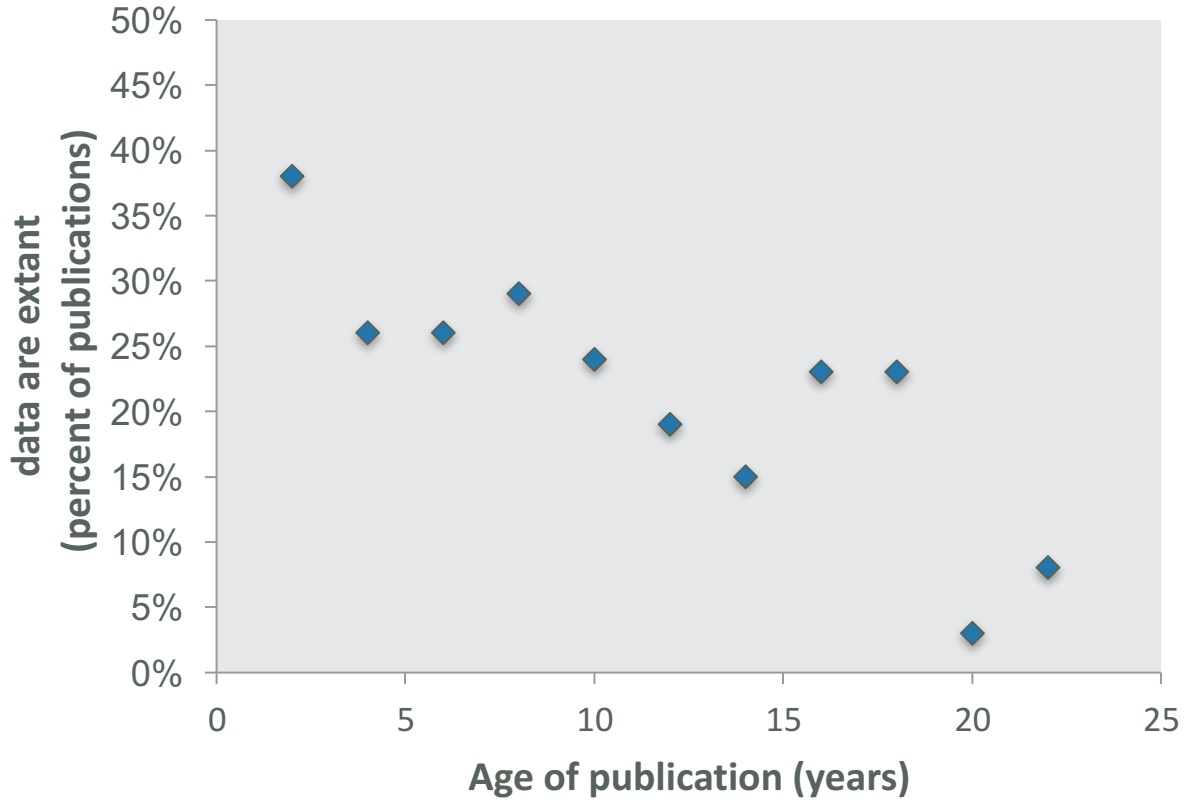
Fundamental sharing policy for *Nature* and the Nature research journals

> An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature journal is that **authors are required to make materials, data, code, and associated protocols promptly available** to readers without undue qualifications. Any restrictions on the availability of materials or information must be disclosed to the editors … [and] … in the submitted manuscript.

Supporting data must be made available to editors and peer-reviewers at the time of submission for the purposes of evaluating the manuscript.

See http://www.nature.com/authors/policies/availability.html

Sharing upon request has problems



Replotted from:
Vines *et al*. *Current Biology* (2014)
doi:10.1016/j.cub.2013.11.014

Raw data at *Dryad*
doi: 10.5061/dryad.q3g37

Data-access practices strengthened in Nature journals

- Clear preference for sharing large datasets via public repositories.
- Enforce data deposition in fields where there is strong community consensus
- List of public data repositories now maintained by *Scientific Data*
- Encourage authors to publish Data Descriptors at *Scientific Data*
  - before, with or after the analysis paper
  - editors work with authors
- Data availability statements required

SCIENTIFIC DATA

Share your data in a useful manner

3

nature research

How do you make your data useful?

Open data is about more than disclosure – it must be "FAIR"

- Findable
- Accessible
- Interoperable
- Re-usable

Wilkinson *et al. Sci. Data* **doi:10.1038/sdata.2016.18** (2016)

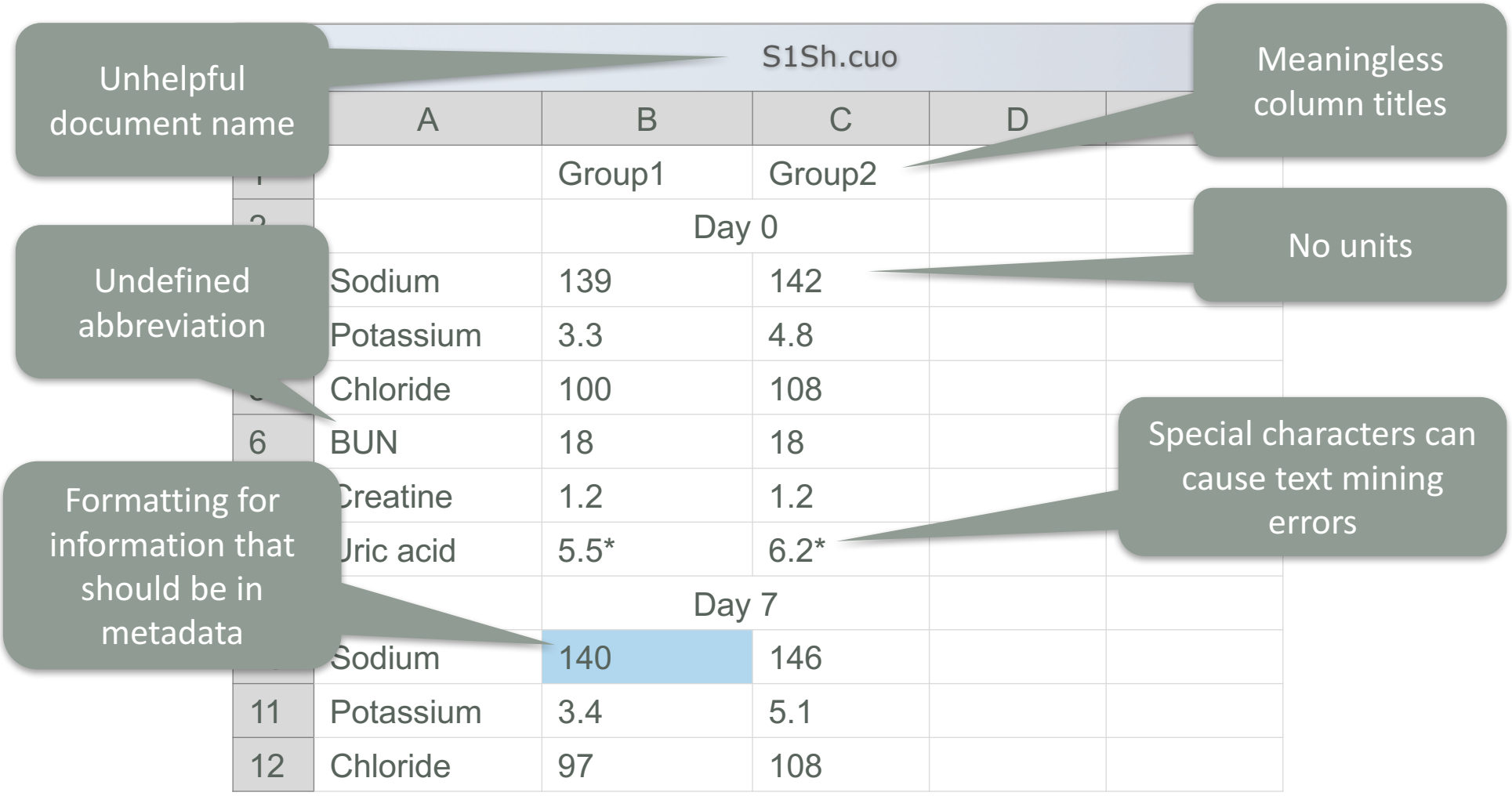
Publish it as supplementary information

- Often sufficient for small datasets
- Generally stable over time
- Generally compatible with a range of information types

But,

- Often poorly curated and not always thoroughly reviewed – quality varies significantly
- Not often machine-readable
- Hard to cite separately, lack of credit

Poor data layout



S1Sh.cuo

	A	B	C	D
1		Group1	Group2	
2		Day 0		
3	Sodium	139	142	
4	Potassium	3.3	4.8	
5	Chloride	100	108	
6	BUN	18	18	
7	Creatine	1.2	1.2	
8	Uric acid	5.5*	6.2*	
9		Day 7		
10	Sodium	140	146	
11	Potassium	3.4	5.1	
12	Chloride	97	108	

Better data layout

Table_S1_Shanghai_blood.xls

	A	B	C	D	E	F
1	Parameter	Day	Control	Treated	Units	P
2	Sodium	0	139	142	mEq/l	0.82
3	Sodium	7	140	146	mEq/l	0.70
4	Sodium	14	140	158	mEq/l	0.03
5	Sodium	21	143	160	mEq/l	0.02
6	Potassium	0	3.3	4.8	mEq/l	0.06
7	Potassium	7	3.4	5.1	mEq/l	0.07
8	Potassium	14	3.7	4.7	mEq/l	0.10
9	Potassium	21	3.1	3.6	mEq/l	0.52
10	Chloride	0	100	108	mEq/l	0.56
11	Chloride	7	97	108	mEq/l	0.68
12	Chloride	14	101	106	mEq/l	0.79

Know the relevant standards in your community

- Many communities have developed specific guidelines for reporting certain kinds of data
- Check journal guidelines to see what is required.
- Can help you format your data in a useful manner

biosharing (https://biosharing.org)

- Browse information on over 600 reporting standards
- Find standards that are relevant to your type of data

Find the right repository for your data

4

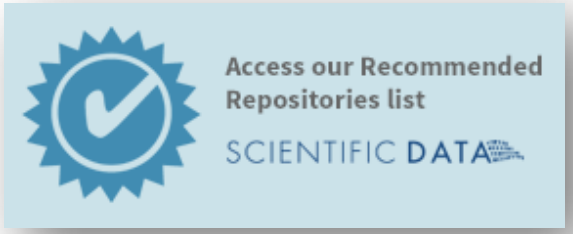
nature research

Find the right repository for your data

What to look for in a data repository

- Quality curation
- A commitment to long-term preservation
- Features that support collaborative analysis
- Features that allow you keep data private until you are ready to publish.
- **Investigate data archiving options at your institution**

Find the right repository for your data



http://www.nature.com/sdata/policies/repositories

Browse our recommended data repository online.

- *We currently list more than 90 repositories, across the biological, physical and social sciences*
- *We advise authors on the best place to store their data*

Publish your data

5

nature research

the data paper

SCIENTIFIC DATA

- A clear, peer reviewed description of data, to maximize usage
- Citable publications that give credit for reusable data
- Several journals publish data paper formats, including *GigaScience*, *Earth Systems Science Data*, and at Nature*, Scientific Data*

nature research

Get Credit for Sharing Your Data

Publications will be indexed and citeable.

Open-access

Articles are published by default under a Creative Commons Attribution licence (CC BY). Each publication supported by CC0 metadata.

Focused on Data Reuse

All the information others need to reuse the data; no interpretative analysis, or hypothesis testing

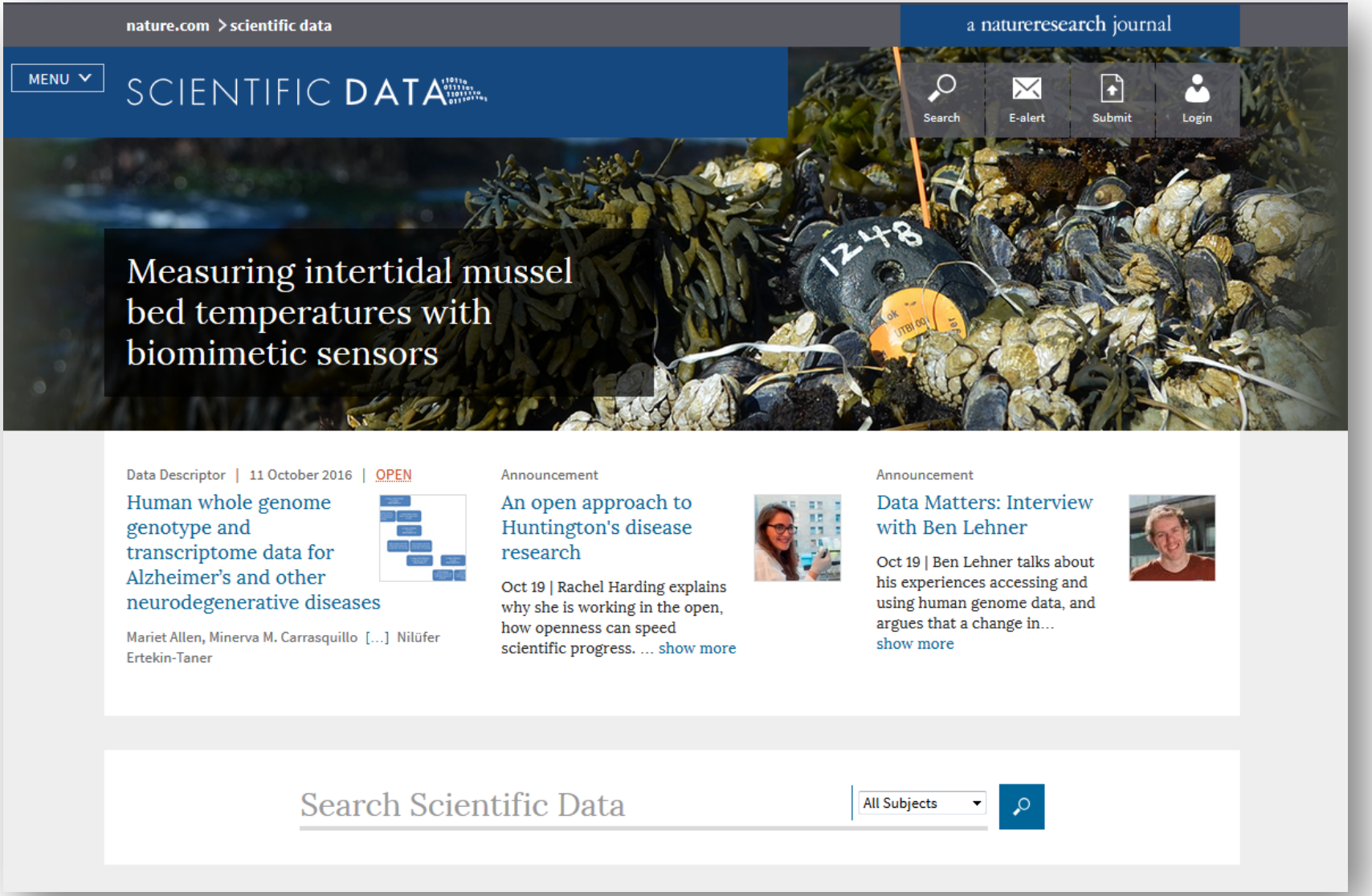
Peer-reviewed

Rigorous peer-review focused on technical data quality and reuse value

Promoting Community Data Repositories

Not a new data repository; data stored in community data repositories

Launched in May 2014



SCIENTIFIC DATA

The Data Descriptor article-type

Data Descriptor

Focus on data reuse

- Detailed descriptions of the methods and technical analyses supporting the quality of the measurements.
- Does not contain tests of new scientific hypotheses

Sections:

- Title
- Abstract
- Background & Summary
- Methods
- Data Records
- Technical Validation
- Usage Notes
- Figures & Tables
- References
- Data Citations

Data Records

All the samples used in this study are summarized in Table 1. Consistent identifiers are used in Tables 2 and 3 to allow mapping between the proteomic and transcriptomic data outputs.

Data Record 1
The raw data, peaklists (.mgf), ProteomeDiscoverer result files (.msf) and ProteomeDiscoverer workflow files (.xml) have been uploaded to ProteomeXchange (http://www.proteomexchange.org/) with the following accession number PXD000134 (ref. 67; Table 2).

Data Record 2
Microarray data are available at the NCBI Gene Expression Omnibus (GEO) database under the accession numbers GSE26451 (ref. 68) and GSE26453 (ref. 69; Table 3).

Data Record 3
The peptide and protein identification data sets have been annotated by The Global Proteome Machine at http://gpmdb.thegpm.org/

Data Record 4
The peptide and protein identification data sets have been annotated by the StemCellOmicsRepository (SCOR) at http://scor.chem.wisc.edu/

Data Citations

67. Low, T. Y. *et al.* ProteomeXchange: PXD000134 (2013).
68. Chin, A. *et al.* Gene Expression Omnibus: GSE26451 (2011).
69. Chin, A. *et al.* Gene Expression Omnibus: GSE26453 (2011).

A Data Descriptor

SCIENTIFIC DATA

SCIENTIFIC DATA

Altmetric: 23 Views: 17,478 Citations: 26 More detail »

PDF Share Share Tools

Data Descriptor | OPEN

Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies

Glenn S Cowley, Barbara A Weir […] William C Hahn

Scientific Data **1**, Article number: 140035 (2014)
doi:10.1038/sdata.2014.35
Download Citation

Cancer genomics RNAi

Received: 20 May 2014
Accepted: 22 August 2014
Published online: 30 September 2014

Corrigendum (11 November 2014)

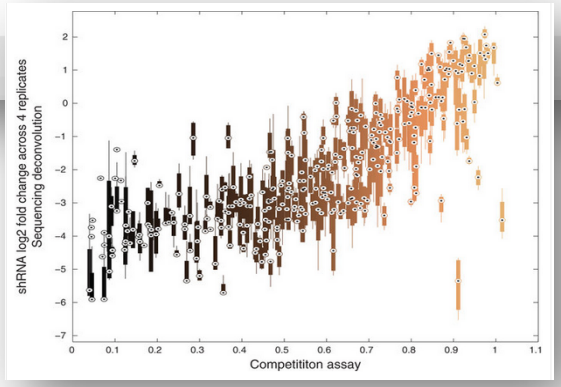
Associated Content

Cancer Discovery | Article
Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells
A. Buzina, A. Datti […] B. G. Neel

Proceedings of the National Academy of Sciences | Article
Highly parallel identification of essential genes in cancer cells
A. Subramanian, B. A. Weir […] C. Li

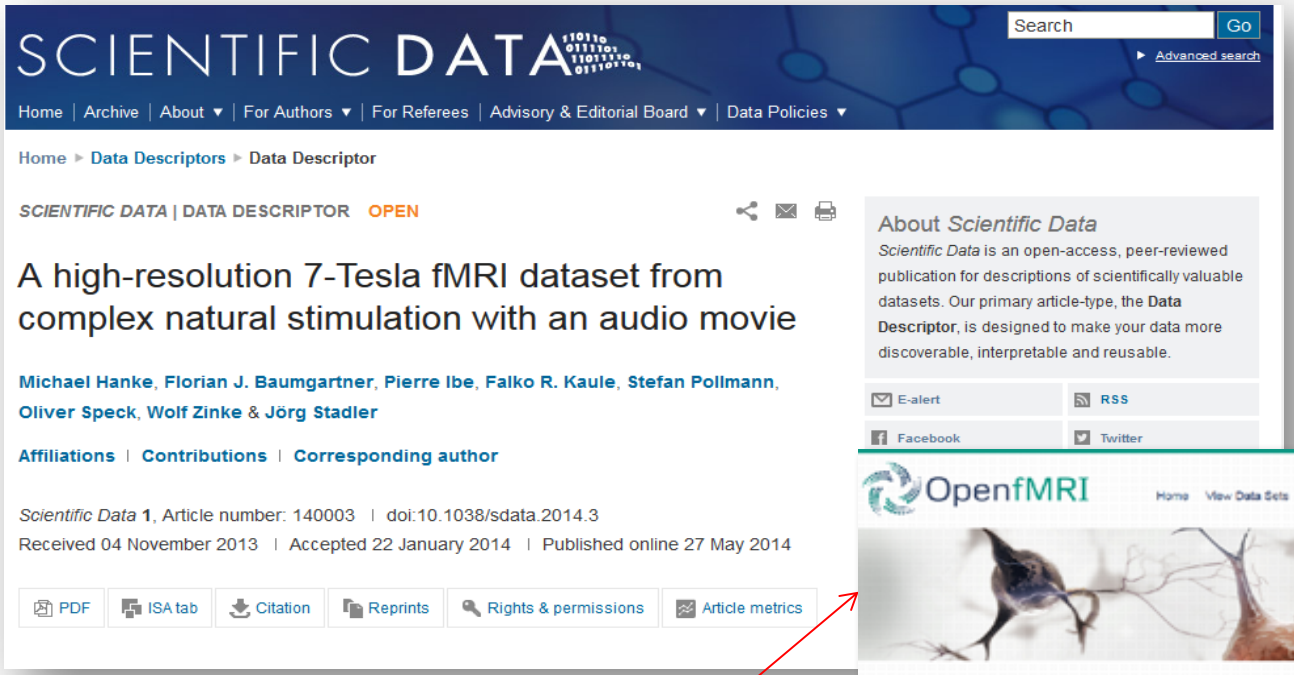
Proceedings of the National Academy of Sciences | Article
Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer
A. East, A. Tsherniak […] C. H. Mermel

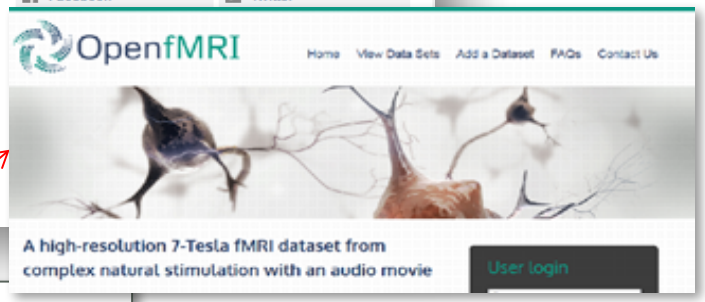
Abstract



- **Screen results and in-depth analysis published in 2011 at *PNAS***
- **Full screen data published at *Scientific Data* in 2014**
- Data at **figshare**
- Data Descriptor cited 89 times according to Google Scholar!

Neuroscience





- **Previously unpublished dataset**
- Data in OpenfMRI
- Cited 39 times according to Google Scholar
- Authors ran an analysis challenge after publication

Think beyond your own data

cite the data of others when you use it

be a data-minded peer-reviewer

use data responsibly

Get the most from your data

Preserve it
Encourage reuse
Get credit

Encourage others to do the same

SCIENTIFIC DATA

Thanks!

Managing Editor
Andrew L. Hufton
andrew.hufton@nature.com

Honorary Academic Editor
Susanna-Assunta Sansone

Data Curation Editor
Varsha Khodiyar

Senior Publishing Assistant
Joseph Salter

Head of Data Publishing
Iain Hrynaszkiewicz

Visit nature.com/scientificdata

Email scientificdata@nature.com

Tweet @ScientificData

Supported by DRYAD biosharing figshare credit for all your research isatools COLLECT | CURATE | ANALYSE | SHARE | PUBLISH

nature research