

Making data publication a first class research output

Andrew L. Hufton

Managing Editor, *Scientific Data*

<https://www.nature.com/sdata/>

Helping Researchers Publish,

University of Cambridge,

Oct 2017

Launched in May 2014

nature.com > scientific data

a natureresearch journal

MENU


SCIENTIFIC DATA

Search E-alert Submit Login

Measuring intertidal mussel bed temperatures with biomimetic sensors

Data Descriptor | 11 October 2016 | [OPEN](#)

Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases




Mariet Allen, Minerva M. Carrasquillo [...] Nilüfer Ertekin-Taner

Announcement

An open approach to Huntington's disease research


Oct 19 | Rachel Harding explains why she is working in the open, how openness can speed scientific progress. ... [show more](#)



Announcement

Data Matters: Interview with Ben Lehner

Oct 19 | Ben Lehner talks about his experiences accessing and using human genome data, and argues that a change in... [show more](#)



Search Scientific Data

All Subjects

the data paper

SCIENTIFIC DATA 



- A clear, peer reviewed description of data, to maximize usage
- Citable publications that give credit for reusable data
- Visit our journal homepage to learn more: <https://www.nature.com/sdata/>



Get Credit for Sharing Your Data

Publications will be indexed and citeable.



Open-access

Data Descriptors are published 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

The Data Descriptor article-type

1

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).

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](https://doi.org/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](#)

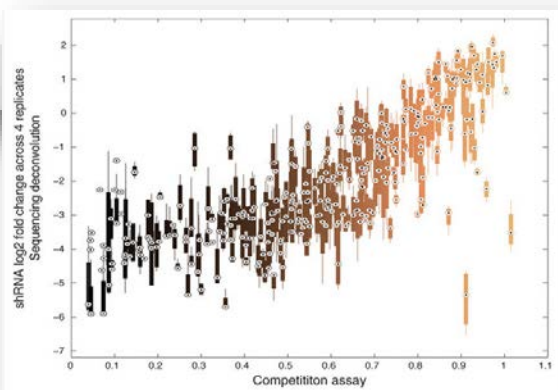
A. Subramanian, B. A. Weir [...] C. Li

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 117 times according to Google Scholar!

SCIENTIFIC DATA

Altmetric: 57 Citations: 27 [More detail >>](#)

Data Descriptor | [OPEN](#)

A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie

Michael Hanke , Florian J. Baumgartner, Pierre Ibe, Falko R. Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke & Jörg Stadler

Scientific Data **1**, Article number: 140003 (2014)
 doi:10.1038/sdata.2014.3

Received: 04 November 2013
 Accepted: 22 January 2014
 Published online: 27 May 2014

Data Citations

[Abstract](#) • [Background & Summary](#) • [Methods](#) • [Data Records](#) • [Technical Validation](#) • [Usage Notes](#) • [Additional information](#) • [References](#) • [Data Citations](#) • [Acknowledgements](#) • [Author information](#) • [Supplementary information](#)

- Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. *OpenfMRI ds000113* (2014).

- **Previously unpublished dataset**
- Data in OpenfMRI
- Cited 42 times according to Google Scholar
- Authors ran an analysis challenge after publication

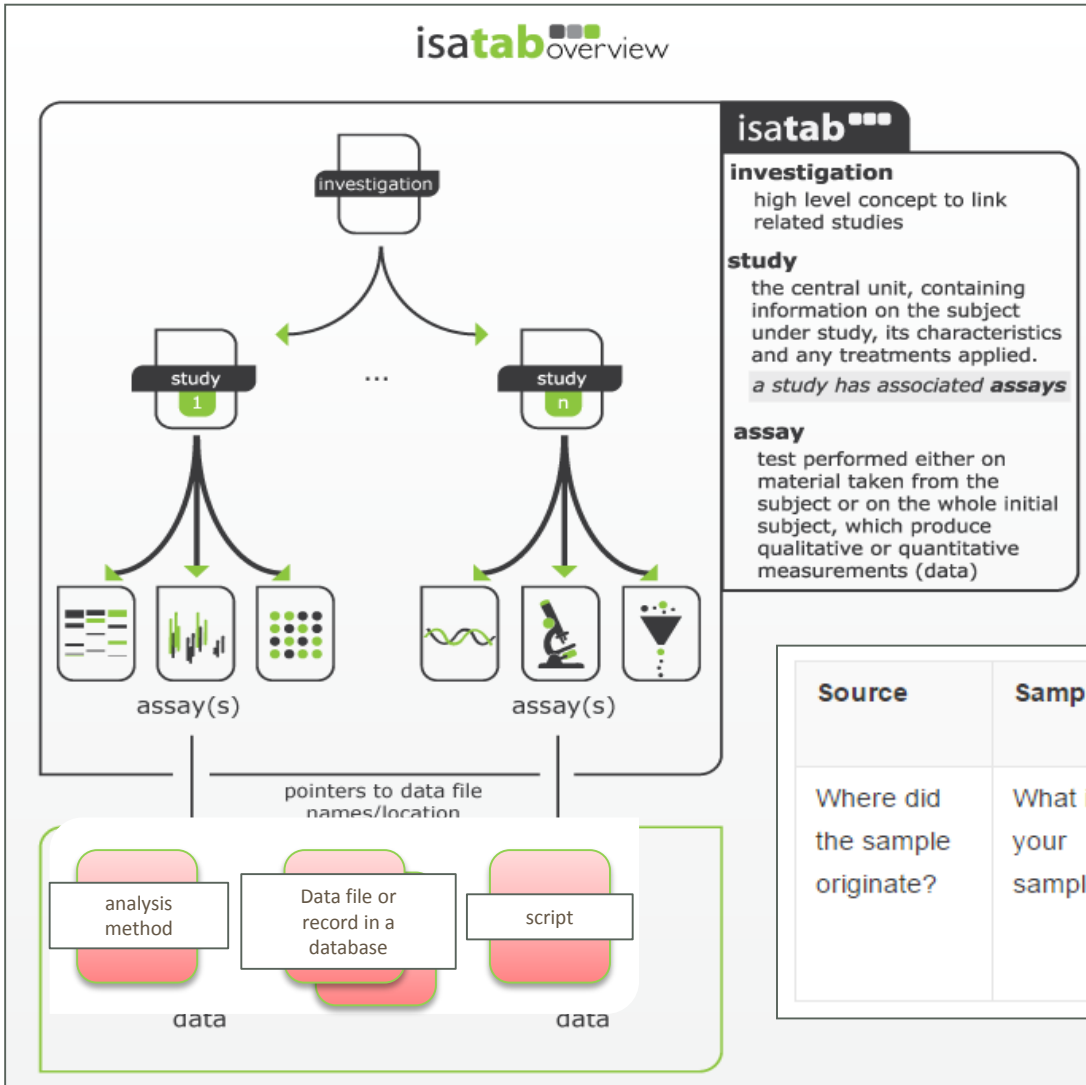


Code in GitHub

- Additional resources**
- More information and updates available at: <http://www.studyforrest.org>
 - Source code repository: <http://github.com/hanke/gumpdata>

Data Descriptor

Structured metadata



In-house curation team:

- assists users to submit the structured content via simple templates and an internal authoring tool
- performs value-added semantic annotation of the experimental metadata

Source	Sample	Characteristics	Protocol	Data
Where did the sample originate?	What is your sample?	What do users need to know about your source and samples to fully understand your work?	How did the samples become data?	Where is the data?

Explore this structured metadata with ISA-explorer

isaexplorer
SCIENTIFIC DATA

Search

Data Repositories

figshare	36
Dryad Digital Repository	32
Gene Expression Omnibus	17
ProteomeXchange	5
NCBI Sequence Read Archive	5
1000 Functional Connectomes Project International Neuroimaging Data-Sharing Initiative (FCP/INDI)	5
Harvard Dataverse Network	5
MetaboLights	4

[Show all](#) [Show next 5](#) [Reset](#)

Designs

124 Data Descriptor Articles Displayed

12/11/2013

Zengchao Hao et al

Global integrated drought monitoring and prediction system

Data Repositories **1**

WATER RESOURCES HYDROLOGY

26/02/2014

Amelie Baud et al

Genomes and phenomes of a population of outbred rats and its progenitors

Data Repositories **10**

GENETIC VARIATION
GENOME-WIDE ASSOCIATION STUDIES

<http://scientificdata.isa-explorer.org/>

credit Alejandra Gonzalez-Beltran, Susanna Assunta-Sansone

nature research

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)

<https://www.nature.com/articles/sdata201618>

Helping authors find the right repository for their data

2

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*
- *Support the use of institutional repositories, including **the University of Cambridge data repository** (<https://www.data.cam.ac.uk/repository>)*

A rigorous, community-based editorial process

3

Criteria for publication

<https://www.nature.com/sdata/policies/for-referees>

- Acceptance for publication is based on the
 - technical rigour of the procedures used to generate the data
 - the reuse value of the data
 - the completeness of the data description
- Evaluation is not be based on the perceived impact or novelty of the findings associated with the datasets.

Data policies

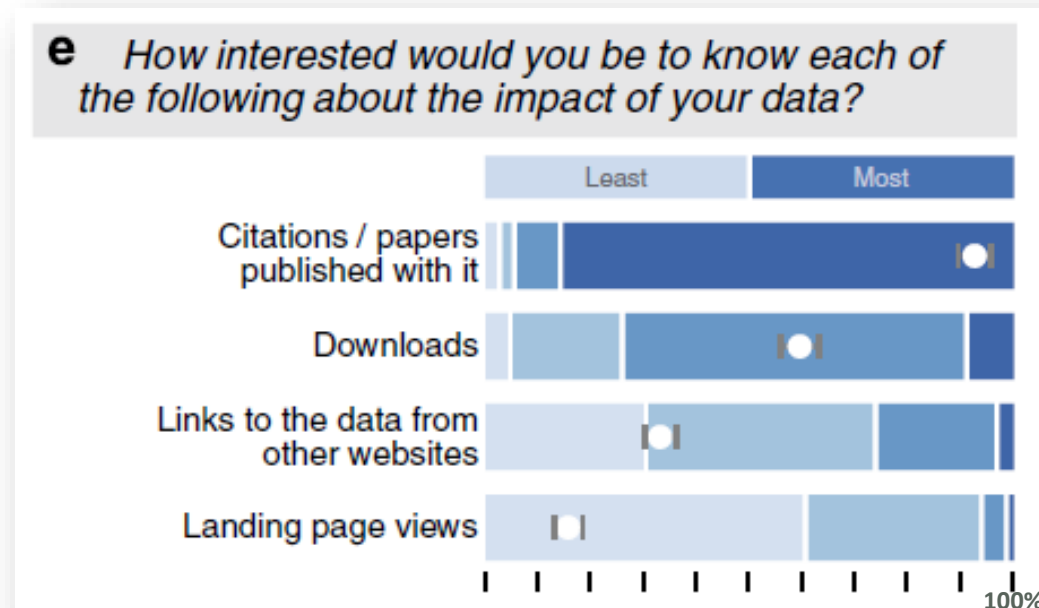
<https://www.nature.com/sdata/policies/data-policies>

- **Editors and peer-reviewers have the right to view the data during their evaluation**
 - Before peer-view we check:
 - Links to the datasets work (see first page of PDF)
 - Data are hosted in a repository that meets our policies
- **At time of publication, authors must publicly release the data under terms that permit wide reuse (i.e. as “open data”)**
 - Restrictions on commercial use or redistribution are not permitted

Promoting a culture of credit for data sharing

4

Researchers want to know when their data is used in published works



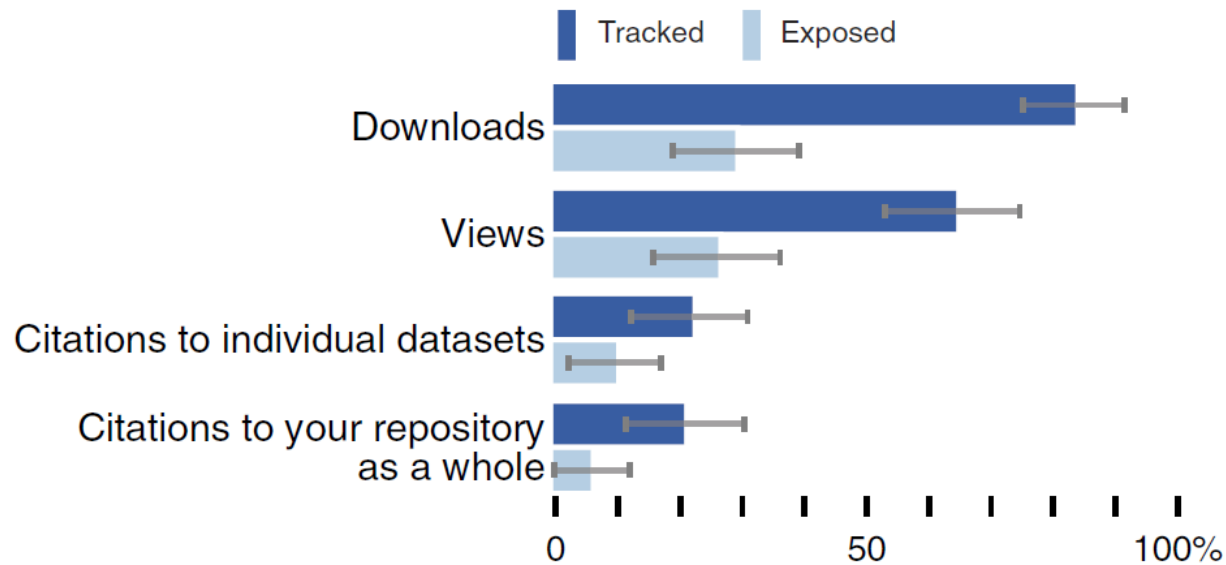
From

Making data count

John E. Kratz & Carly Strasser. *Sci. Data*
doi:10.1038/sdata.2015.39 (2015)

But repositories have a hard time tracking actual data use

f *What metrics/statistics do your repository currently track and expose?*



From

Making data count

John E. Kratz & Carly Strasser. *Sci. Data*
doi:10.1038/sdata.2015.39 (2015)

In-article data citation

SCIENTIFIC DATA | DATA DESCRIPTOR OPEN



Plant traits, productivity, biomass and soil properties from forest sites in the Pacific Northwest, 1999–2014

The dataset (*NACP TERRA-PNW: Forest Plant Traits, NPP, Biomass, and Soil Properties, 1999–2014*) is hosted with other contributions from the North American Carbon Program (NACP) by the Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics
 (Data Citation 1) Oak Ridge National Laboratory Distributed Active Archive Center

2016



PDF



ISA tab



Data Citations

[Abstract](#) • [Background & Summary](#) • [Methods](#) • [Data Records](#) • [Technical Validation](#) • [Additional Information](#) • [References](#) • [Data Citations](#) • [Acknowledgements](#) • [Author information](#)

1. Law, B. E., & Berner, L. T. *Oak Ridge National Laboratory Distributed Active Archive Center*
<http://dx.doi.org/10.3334/ORNLDAAC/1292> (2015).

Abstract

[Abstract](#) • [Background & Summary](#)
 • [References](#) • [Data Citations](#)

Plant trait measurements are needed for evaluating ecological responses to environmental

Get the most from your data

Preserve it
Encourage reuse
Get credit

Encourage others to do the same

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 Hrynaskiewicz

Visit nature.com/scientificdata

Email scientificdata@nature.com

Tweet [@ScientificData](https://twitter.com/ScientificData)

Supported by

