THE ROLE OF PUBLISHERS IN REPRODUCIBLE RESEARCH

STM Beyond Open Access seminar, December 5th 2014

Iain Hrynaszkiewicz
Head of Data and HSS Publishing, Open Research
Nature Publishing Group & Palgrave Macmillan

iain.hrynaszkiewicz@nature.com
@iainh_z

macmillan
Science and Education

Transforming Learning and Discovery

Beyond open access: Reproducibility

- Open ~~access~~ is a means not the end
- Open access (to papers) is just part of the solution
- Also need code, data, protocols – and research reported in sufficient detail to enable others to understand and repeat

Irreproducibility: underlying issues

(Mis)conduct

Publication bias and refutations – where?

Experimental design

Statistics

Lab supervision and training

Pressure to publish

Reporting and sharing information

Gels, microscopy images

Animal studies description

Methods description

Data deposition

Irreproducibility: all sciences

Miguel et al. (2014). **Promoting transparency in social science research**. *Science (New York, N.Y.)*, *343*(6166), 30–1. doi:10.1126/science.1245317

Recommendations include:

- Preregistration of studies
- Better reporting guidelines
- Sharing of data

Reproducibility: role of publishers

"*Scholarly publishers have an important role in encouraging and mandating the availability of data and...developing innovative mechanisms and platforms for sharing and publishing products of research*"

-- Hrynaszkiewicz I, Li P, Edmunds SC: **Open science and the role of publishers in reproducible research**. In: *Implementing Reproducible Research.* Edited by Stodden V, Leisch F, Peng RD. Chapman & Hall/CRC; 2014

Reproducibility: role of publishers

- Content
- Policies
- Incentives
- Licenses
- Access
- Reliability

Image credit: DS Pugh [CC-BY-SA-2.0 (http://creativecommons.org/licenses/by-sa/2.0)], via Wikimedia Commons. http://commons.wikimedia.org/wiki/File%3AHarlow_Carr_-_geograph.org.uk_-_32309.jpg

Reproducibility: role of publishers

- Content
- Policies
- Incentives
- Licenses
- Access
- Reliability

Image credit: DS Pugh [CC-BY-SA-2.0 (http://creativecommons.org/licenses/by-sa/2.0)], via Wikimedia Commons. http://commons.wikimedia.org/wiki/File%3AHarlow_Carr_-_geograph.org.uk_-_32309.jpg

Reproducibility: Content

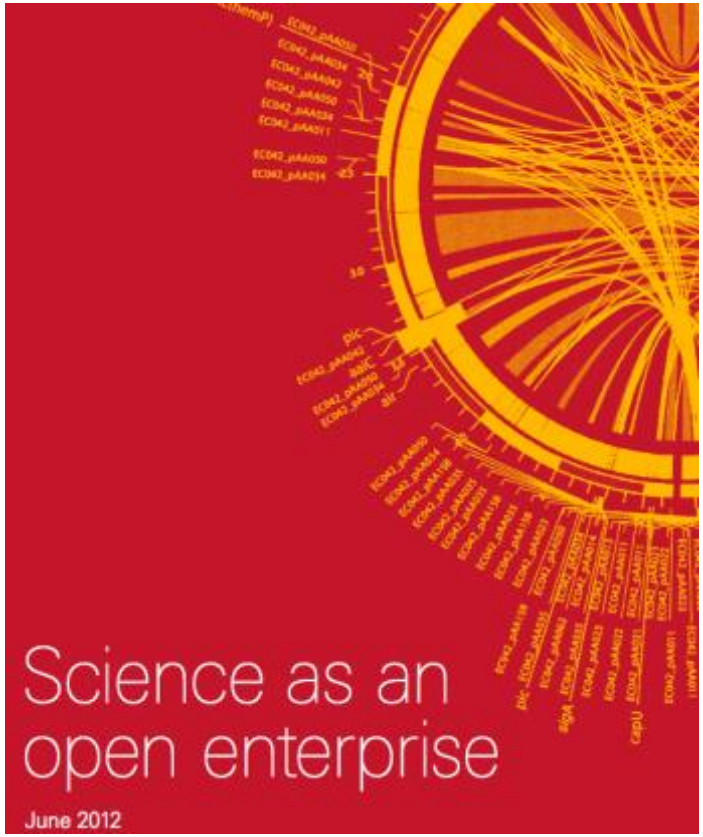
- **Format**
  - Open, standardized XML for articles in PubMed Central; optimal files types for data
- **Amount**
  - Methodological detail and transparent reporting
- **Types**
  - Methods, protocols, data and software papers/journals, short reports, extended reports, updates etc

Reproducibility: Content - examples

- Removal of limitations on Methods sections at Nature journals
- Paul Glasziou (2008) BMJ – inadequate methods descriptions for medical interventions http://www.bmj.com/content/336/7659/1472
- New types of journal and publication…

Role of data journals/articles

- Credit
- Unpublished data
- Peer review focus
- Value of data vs. analysis
- Discoverability
- Reusability
- Narrative/context
- “Intelligently open data”



Data, data (journals) everywhere?

(GIGA)n SCIENCE

F1000 Research

open health data

BMC Research Notes

DATA IN BRIEF
MAKE YOUR DATA COUNT

A peer-reviewed open-access journal
Biodiversity Data Journal
Making your data count! ISSN 1314-2828 (online)

Scientific Data

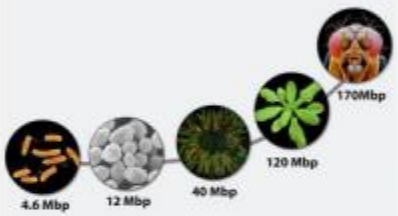
SCIENTIFIC DATA

Search Go

Advanced search

Home | Archive | About ▾ | For Authors ▾ | For Referees | Data Policies ▾ | Collections ▾

Featured Data Descriptor



Long-read, whole-genome shotgun sequence data for five model organisms

Kim *et al.* | 25th November 2014

Third-generation sequencing technologies provide data with unique characteristics including very long read lengths. Here, the authors share high-quality genomic sequencing data from the Pacific Biosciences platform for organisms with storied histories in biological research, ranging from the simple to the complex.

About *Scientific Data*

Scientific Data is an open-access, peer-reviewed publication for descriptions of scientifically valuable datasets. Our primary article-type, the **Data Descriptor**, is designed to make your data more discoverable, interpretable and reusable.

E-alert | RSS

Facebook | Twitter

Submit manuscript

Latest content

Data Descriptor | 25 November 2014

Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction

Matthew D Luciw, Ewa Jarocka & Benoni B Edin

Data Descriptor | 25 November 2014

Simplified data access on human skeletal muscle transcriptome responses to differentiated exercise

Kristian Viasing & Peter Schjerling

Data Descriptor | 25 November 2014

Long-read, whole-genome shotgun sequence data for five model organisms

Kristi E Kim, Paul Peluso [...] Jane M Landolin

Data Descriptor | 11 November 2014

DNA methylation temporal profiling following peripheral versus central nervous system axotomy

Ricco Lindner, Radhika Puttagunta [...] Simone Di

Get Credit for Sharing Your Data

Publications will be indexed and citeable.

Open-access

Creative Commons licenses (CC-BY/CC-BY-NC) for the main Data Descriptor. Each publication supported by CCO 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

Neuroscience

SCIENTIFIC DATA

- **New Dataset**
- Data in OpenfMRI
- Source code in GitHub
- *Big Data*

SCIENTIFIC DATA

Search Go

Advanced search

Home | Archive | About ▾ | For Authors ▾ | For Referees | Advisory & Editorial Board ▾ | Data Policies ▾

Home ▸ Data Descriptors ▸ Data Descriptor

SCIENTIFIC DATA | 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 & Jorg Stadler

Affiliations | Contributions | Corresponding author

Scientific Data 1, Article number: 140003 | doi:10.1038/sdata.2014.3

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

PDF | ISA tab | Citation | Reprints | Rights & permissions | Article metrics

About *Scientific Data*

Scientific Data is an open-access, peer-reviewed publication for descriptions of scientifically valuable datasets. Our primary article-type, the **Data Descriptor**, is designed to make your data more discoverable, interpretable and reusable.

E-alert | RSS | Facebook | Twitter

OpenfMRI

Home | View Data Sets | Add a Dataset | FAQs | Contact Us

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

User login

Data Citations

Abstract • Background & Summary • Methods • Data Records • Technical Validation • Usage Notes • Additional information • References • Data Citations • Acknowledgements • Author information • Supplementary information

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

Additional resources:

- More information and updates are made availa at: http://www.studyforrest.org
- Source code repository: http://github.com/hanke/gumpdata
- Documentation for the source code: http://gumpdata.readthedocs.org

Code in GitHub

nature publishing group npg

Reproducibility: Policies

- **Adherence to reporting and minimum information standards**
  - Checklists and enforcement
- **Data deposition**
- **Data and materials sharing**
- **Encouraging better practice**
  - Encourage publication of Data Descriptors

Mandates aren't always enough

2002: Nature journals mandate deposition of MIAME-compliant microarray data
2006: compliance issues identified

Ioannidis *et al.*, Nat Gen 41, 2, 149 (2009)

Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

Of 18 papers published in Nat Gen in 2005-2006, 10 analyses could not be reproduced, 6 only partially.

2013

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Arch

Archive > Volume 496 > Issue 7446 > Editorial > Article

NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

nature medicine

nature.com ▸ journal home ▸ archive ▸ issue ▸ editorial ▸ full text

NATURE MEDICINE | EDITORIAL

日本語要約

Raising stan

EDITORIAL

nature structural & molecular biology

Raising standards

Nature journals' updated editorial policies aim to improve transparency and reproducibility.

nature immunology

Raising standards

NATURE CHEMICAL BIOLOGY | EDITORIAL

Facilitating reproducibility

nature cell biology

Raising reporting standards

Nature journals' updated editorial policies aim to improve trans

EDITORIAL

Raising standards

nature biotechnology

Nature Biotechnology and other Nature journals are updating editorial policies with the aim of improving transparency and reproducibility.

nature neuroscience

Raising standards

Nature journals' updated editorial policies aim to imp

nature genetics

Raising standards

Enhancing reproducibility

NATURE METHODS | VOL.10 NO.5 | MAY 2013 | 367

Reproducibility: Policies

ANNOUNCEMENT

Data-access practices strengthened

In our continued drive for reproducibility, *Nature* and the Nature research journals are strengthening our editorial links with the journal *Scientific Data* and enhancing our data-availability practices. We believe that this initiative will improve support for authors looking for appropriate public repositories for their research data, and will increase the availability of information needed for the reuse and validation of those data.

In 2013, Nature journals introduced new editorial measures to promote reproducibility, and we continue to evaluate their impact and refine our policies. Our newly strengthened data-availability practices (go.nature.com/o5ykhe) reflect our preference that data be deposited in public repositories, and encourage researchers to expand on work published in the Nature journals by publishing further information in *Scientific Data*.

Community-supported, specialized data repositories are usually the best way to share large data sets. General, unstructured repositories, such as figshare and Dryad, provide options where no community repository exists, and are preferable to publishing data as Supplementary Information. Supplementary materials have size limitations and do not always provide optimal file and viewing formats, particularly for large and complex data sets. But where no repository — or publication focused on detailed descriptions of data sets — exists, supplementary materials have often been the best option.

Scientific Data (go.nature.com/iyu9qh), which launched this year, offers authors another way to maximize the value of their data sets for further research — for themselves and for the scientific community.

Its primary article type, the Data Descriptor, provides more detail to improve the data's discoverability, interpretability and reusability — as well as allowing the highest credit to be given to the authors who created the data set.

We are now rolling out a new process under which, when they accept a manuscript containing appropriate data sets, editors of *Nature* and Nature research journals will encourage authors to submit the data sets to *Scientific Data* as a Data Descriptor (go.nature.com/utfvfo).

Authors may also submit a Data Descriptor manuscript alongside a manuscript for a Nature journal. If appropriate, they could publish the descriptor first, without compromising the novelty of future primary-research articles based on the data. In these cases, authors are encouraged to consult with the editor of their target journal to ensure that prior publication of a Data Descriptor is acceptable. (Note that other publishers may have different policies.)

Scientific Data's peer-review and in-house curation processes focus on ease of reuse. A data-curation editor reviews data files, checks their format, archiving and annotations, and works with authors to produce a standardized, machine-readable summary of the study in the ISA-Tab format (S. Sansone *et al. Nature Genet.* **44,** 121–126; 2012).

Data Descriptors can accommodate all data types, including raw data and updated data sets generated after initial publication. They can also show the controls required for validation of the data set, which may have been excluded from the primary paper because of space limitations. *Scientific Data*'s editorial process assesses repositories and helps to ensure that data are placed in the correct one. *Nature*'s enhanced data-availability policy now directs authors to a list of approved repositories (go.nature.com/jpm768).

Several articles published in Nature research journals already have complementary articles in *Scientific Data* (such as A. Baud *et al. Sci. Data* **1,** 140011 (2014) and F. Roquet *et al. Sci. Data* **1,** 140028; 2014). As science evolves and produces ever-increasing amounts of data, those data must be collected, organized, curated, quality-checked and made available on the right platform so that they can be easily discovered and reused. Stronger links with *Scientific Data* and our data-availability practices aim to achieve this. ■

312 | NATURE | VOL 515 | 20 NOVEMBER 2014

2014

Reproducibility: Data policy examples

- Data sharing statements in published papers (*Annals Internal Medicine*, BMJ [non-clinical trials]*)
- Data sharing implied by submission (BMC min. requirement)
- Data sharing as a condition of publication (Nature journals min. requirement)
- Mandated data sharing as a condition of publication (PLOS)
- Mandated data sharing with statement and link in article (ecology journals signed up to joint data archiving policy)
- Mandated open data as a condition of submission (*Scientific Data, F1000Research*)

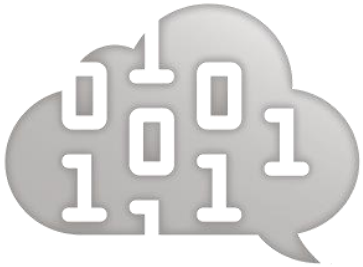
Reproducibility: Incentives

- Data and code citation
- Data articles and journals
- Recognising reproducibility – challenges, awards
- Demonstrating impact

Reproducibility: Licenses

Articles: Creative Commons licenses

Metadata: released under the **CC0 waiver** to maximize reuse and aid data miners



Data: depends on public repositories. Some repositories e.g. figshare and Dryad both use the **CC0 waiver**.

Reproducibility: Access

- Discoverability and links to other digital products of research
- Links between papers
  - Nature ENCODE Explorer
  - Threaded publications (BMC/Crossref/Others)
- Repository partnerships
- Integration with tools e.g. document authoring, data management, Lab notebooks

Reproducibility: Reliability

- Peer review and editorial process – focus on reproducibility
- Correcting the record
- Evaluating effectiveness of policies

Reproducibility: Reliability - example

Peer review at *Scientific Data* focuses on:

- Completeness (can others reproduce?)
- Consistency (were community standards followed?)
- Integrity (are data in the best repository?)
- Experimental rigour and technical quality (were the methods sound?)

Does not focus on:

- Perceived impact/importance
- Size/complexity of data

Implementation of Nature checklist

Onerous:

Authors, referees, editors, copyeditors

Referees:

We are not yet sure whether they are paying much attention

Authors:

Some papers submitted with checklist without prompt

Many have embraced source data

Improves reporting

We have commissioned an external assessment of the impact

The list may be driving changes in experimental design


Design study
Conduct study
Analyse and deposit data
Write manuscript
Submission
Peer review
Publication
New hypothesis
Publisher

Thank you

For more information
please contact

IAIN HRYNASZKIEWICZ
Head of Data and HSS Publishing,
Open Research

M: +44 (0)7814 290576
T: +44 (0)207 0146753
E: iain.hrynaszkiewicz@nature.com