

# Publication and access to research datasets

UCL & Yale workshop, 9<sup>th</sup> October 2015

**Iain Hrynaszkiewicz**

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

[iain.hrynaszkiewicz@nature.com](mailto:iain.hrynaszkiewicz@nature.com)

[@iainh\\_z](https://twitter.com/iainh_z)



# Publishers/journals and data access

---

- More reliable evidence – and papers
- Journal mission/goals
- Content innovation (and more use and reuse)
- Reliability (peer review)
- Discoverability and visibility (bibliographic databases)
- Permanence (content and links)
- Credit/incentives (article types and citations)
- Encouraging and implementing good practice and policies

# Journal data policies

---

- Willingness to share stated (*Annals Internal Medicine*)
- Data sharing implied by submission (BioMed Central\*)
- Data sharing implied as a condition of publication (Nature\*)
- Mandated data sharing with statement in paper (PLOS, BMJ)
- Mandated data sharing with statement and link to data (non-medical journals e.g. ecology, animal genomics)
- Mandated open data as a condition of submission (*Scientific Data, GigaScience, F1000Research*)



\*Minimum publisher requirement – some disciplines/journals may mandate open access to data

**STRONGER**

1. Vines, T. H. *et al.* **Mandated data archiving greatly improves access to research data.** *FASEB J.* fj.12–218164– (2013). doi:10.1096/fj.12-218164

# Data sharing via supplementary files

TRIALS IMPACT FACTOR 2.12

Search Trials for  Go

Advanced search

Home Articles Authors Reviewers About this journal My Trials

Top  
Abstract  
Background  
Methods  
Results  
Discussion  
Competing interests  
Note for user...  
Authors' contributions  
Sources of Funding  
Acknowledgements  
References

This article is part of the series [Sharing clinical research data](#).

This article is part of the series [F](#)

A [correction](#) for this article has been published.

**Research**

**The International Stroke Trial Collaborative Group**

Peter AG Sandercock<sup>1\*</sup>, Maciej Niewada<sup>2</sup>, Aneta Sandercock<sup>3</sup>

\* Corresponding author: Peter AG Sandercock [Peter.Sandercock@ed.ac.uk](mailto:Peter.Sandercock@ed.ac.uk)

1 Department of Clinical Neurosciences, Western General Hospital, Edinburgh, UK

2 Department of Clinical and Experimental Neurosciences, Krakowskie Przedmieście 26/28, Warsaw, Poland

3 2nd Department of Neurology, Warsaw, Poland

Email: Peter AG Sandercock [Peter.Sandercock@ed.ac.uk](mailto:Peter.Sandercock@ed.ac.uk); Maciej Niewada [maciej.niewada@wum.edu.pl](mailto:maciej.niewada@wum.edu.pl); Aneta Sandercock

*Trials* 2011, **12**:101 doi:10.1186/1745-6215-12-101

The electronic version of this article is the complete one and can be found online at: <http://www.trialsjournal.com/content/12/1/101>

**Results**

Consent for publication of raw data was not obtained from participants. Consent for participation in the trial was obtained from all subjects or from an appropriate proxy, according to the procedures approved by relevant national and local hospital ethics committees (or Institutional Review Boards [IRB]). These patients were treated 15-20 years ago, and many have died. The dataset (see additional file [1](#) - IST\_data.csv) is fully anonymous in a manner that can easily be verified by any user of the dataset. Patients and hospitals are identified only by an anonymous code; there are no identifying data such as name, address or social security numbers; patient age has been rounded to the nearest whole number. In our view, publication of the dataset clearly presents no material risk to confidentiality of study participants.

**Additional file 1. Database with information completed in IST.**  
Format: CSV Size: 4.6MB [Download file](#)

[OPEN DATA](#)

The dataset includes the following baseline data: age, gender, time from onset to randomisation, presence or absence of atrial fibrillation (AF), aspirin administration within 3 days prior to

**Tools**  
[Download references](#)  
[Download XML](#)

Sandercock *et al*: **The International Stroke Trial database.** *Trials* 2011, **12**:101  
doi:10.1186/1745-6215-12-101

# Data sharing via repository links

## Research

### Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence

*BMJ* 2015 ; 351 doi: <http://dx.doi.org/10.1136/bmj.h4320> (Published 16 September 2015)

Cite this as: *BMJ* 2015;351:h4320

[Article](#)[Related content](#)[Metrics](#)[Responses](#)[Peer review](#)

*Joanna Le Noury, research psychologist<sup>1</sup>, John M Nardo, retired clinical assistant professor<sup>2</sup>, David Healy, professor<sup>1</sup>, Jon Jureidini, clinical professor<sup>3</sup>, Melissa Raven, postdoctoral fellow<sup>3</sup>, Catalin Tufanaru, research associate<sup>4</sup>, Elia Abi-Jaoude, staff psychiatrist<sup>5</sup>*

[Author affiliations](#) ▼

Correspondence to: J Jureidini [Jon.Jureidini@adelaide.edu.au](mailto:Jon.Jureidini@adelaide.edu.au)

**Accepted** 3 August 2015



# Data sharing via repository links

Research

## Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence

BMJ 2015 ;  
Cite this as: B

planned (and, if relevant, registered) have been explained.

Article

Data sharing: Clinical study reports, detailed data tables, and programming code are available on the Dryad Digital Repository (<http://dx.doi.org/10.5061/dryad.bv8j6>) and at [www.Study329.org/](http://www.Study329.org/).

Joanna Le No  
professor<sup>1</sup>, J  
associate<sup>4</sup>, B

Author affili

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

Corresponde

Accepted 3

### References

- 01. ↪ Doshi P, Dickersin K, Healy D, Vedula SS, Jefferson T. Restoring invisible and abandoned trials: a call for people to publish the findings. *BMJ* 2013;346:f2865. [FREE Full Text](#)
- 02. ↪ Keller MB, Ryan ND, Strober M, et al. Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial. *J Am Acad Child Adolesc Psychiatry* 2001;40:762-72. [CrossRef](#) [Medline](#) [Web of Science](#)

03. ↪ McHale J, Lunn J, Lunn J, et al. The impact of the 2007 UK government's 'Freedom of Information' Act on the publication of clinical trial results. *BMJ* 2008;337:a1111-11.



# Data sharing via repository links

Research

Restoring Study 329 in treatment of major depression

BMJ 2015; 351:f2071

Cite this as: Le Nestor E, et al. *BMJ* 2015;351:f2071

[Article](#)

Joanna Le Nestor, professor<sup>1</sup>, Joanna Le Nestor, associate<sup>4</sup>, E. Le Nestor, associate<sup>4</sup>, E. Le Nestor, associate<sup>4</sup>

[Author affiliation](#)

Correspondence to: E. Le Nestor


Accepted 30 October 2015

[References](#)

01. → the first

02. → random



03. →



**DRYAD** About ▾ For researchers ▾ For organizations ▾ C

**Data from: Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence**

**Files in this package**

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the [Dryad Terms of Service](#). To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.  

<b>Title</b>	<b>RIAT Restoring Study 329 data files</b>
<b>Downloaded</b>	28 times
<b>Description</b>	Zip file containing: Original SKB CSR (both Appendix D & Appendix G); ii) RIAT Excel spreadsheet for Harms/Adverse Event data; RIAT Excel spreadsheet for Withdrawals/dropouts from the study. R code used to analyse Efficacy data; Original SKB Study 329 Trial protocol which was followed in Restoring Study 329 (RIAT).
<b>Download</b>	<a href="#">README.txt (767bytes)</a>
<b>Download</b>	<a href="#">Dryad.zip (60.44Mb)</a>
<b>Details</b>	<a href="#">View File Details</a>

Dryad Digital


Commercially, and use is non-

people to publish

ersion: a

[ence](#)

0000



List of potential patient identifiers in datasets	
Identifier (information sources)	Comments
<b>Direct</b>	
Name <sup>8,15</sup>	
Initials <sup>13</sup>	
Address, including full or partial postal code <sup>8,15</sup>	
Telephone or fax numbers or contact information <sup>8,10,12,15</sup>	
Electronic mail addresses <sup>8</sup>	
Unique identifying numbers <sup>8,15</sup>	Generalised HIPAA items 7-10, 18
Vehicle identifiers <sup>8</sup>	
Medical device identifiers <sup>8</sup>	
Web or internet protocol addresses <sup>8</sup>	
Biometric data <sup>8</sup>	
Facial photograph or comparable image <sup>8,10,11,13</sup>	
Audiotapes <sup>11</sup>	
Names of relatives <sup>10</sup>	
Dates related to an individual (including date of birth) <sup>8,9,11,15</sup>	
<b>Indirect—may present a risk if present in combination with others in the list</b>	
Place of treatment or health professional responsible for care <sup>10,15</sup>	Could be inferred from investigator affiliations
Sex <sup>9</sup>	
Rare disease or treatment <sup>10</sup>	
Sensitive data, such as illicit drug use or “risky behaviour” <sup>15</sup>	
Place of birth <sup>10,15</sup>	
Socioeconomic data, such as occupation or place of work, income, or education <sup>9,10,12,15</sup>	MRC requirement is for “rare” occupations only
Household and family composition <sup>15</sup>	
Anthropometry measures <sup>15</sup>	
Multiple pregnancies <sup>15</sup>	
Ethnicity <sup>9</sup>	
Small denominators—population size of <100 <sup>14</sup>	
Very small numerators—event counts of <3 <sup>14</sup>	
Year of birth or age (this article)	Age is potentially identifying if the recruitment period is short and is fully described
Verbatim responses or transcripts <sup>15</sup>	

*“...datasets that contain **three or more indirect identifiers**, such as age or sex, should be reviewed by an independent researcher or ethics committee”*

Hrynaszkiewicz *et al.*, *BMJ* 2010; 340: c181



# Data on (reasonable) request - issues

---

- Meta-analysis fails to launch when <40% IPD available – unanswered requests and refusal to share

*Systematic Reviews* 2014, **3**:97 doi:10.1186/2046-4053-3-97

- Poor availability of psychological research data (only 64/249 datasets available)

*American Psychologist*, Vol 61(7), Oct 2006, 726-728. doi:10.1037/0003-066X.61.7.726

- Data received from 1/10 authors publishing in *PLOS Medicine* and *PLOS Clinical Trials*

PLoS ONE 4(9): e7078. doi:10.1371/journal.pone.0007078

- **Sensitive data repositories (e.g. UKDA)**  
Permanence, curation, persistent identifiers, versioning
- **Data-on-request services (e.g. YODA)**  
Independent governance, scientific review and transparency of access requests, DUAs
  - **Journals/publishers**  
Peer review, visibility, credit/citations, robust links

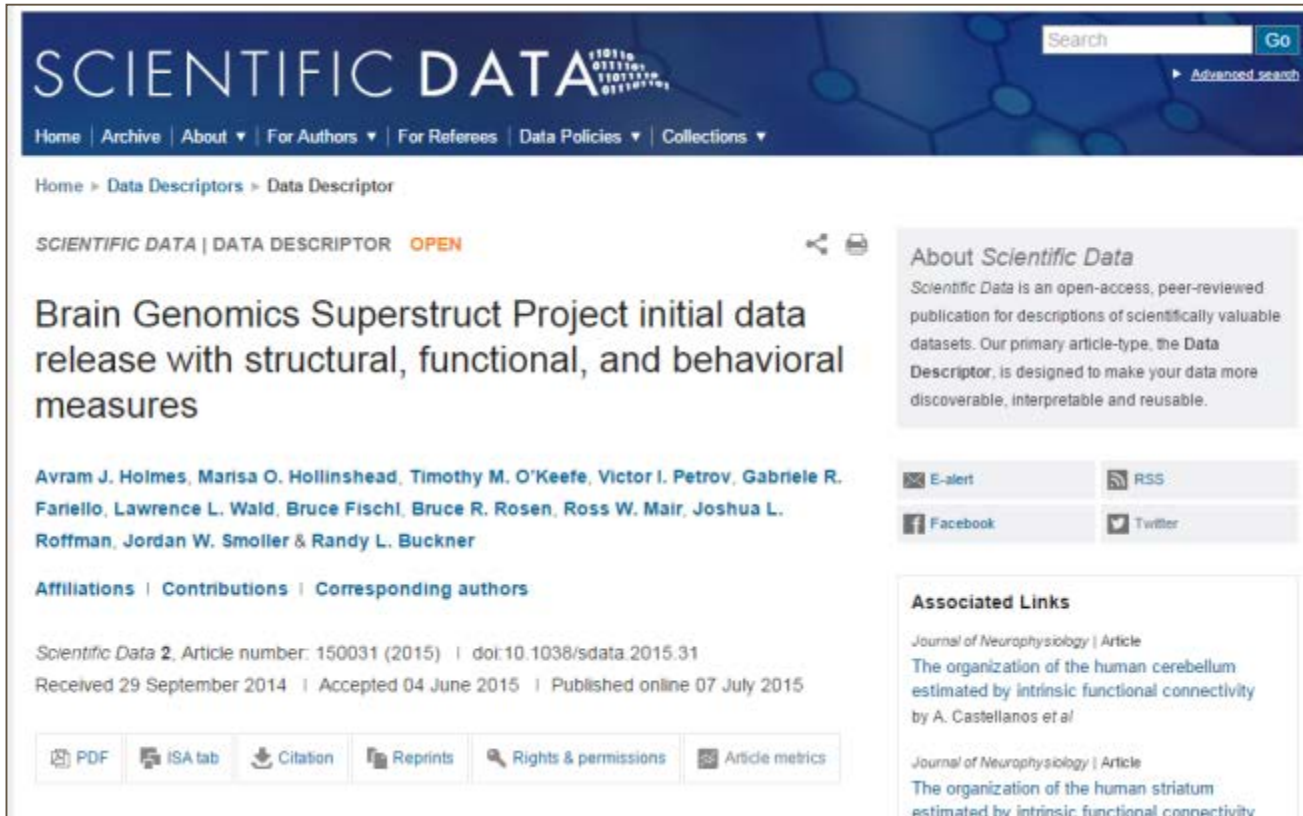


---


## **Better way to *publish* data on request**

1. Hrynaszkiewicz, I., Khodiyar, V., Hufton, A. & Sansone, S. A. **Publishing descriptions of non-public clinical datasets: guidance for researchers, repositories, editors and funding organisations**. BioRxiv <http://dx.doi.org/10.1101/021667> (2015).

# Open access Data Descriptor



The screenshot shows the article page for 'Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures' in Scientific Data. The page includes a search bar, navigation menu, breadcrumb trail, article title, author list, affiliations, publication details, and social media links.

**SCIENTIFIC DATA** 

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

Home > Data Descriptors > Data Descriptor

SCIENTIFIC DATA | DATA DESCRIPTOR **OPEN**

## Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures

Avram J. Holmes, Marisa O. Hollinshead, Timothy M. O'Keefe, Victor I. Petrov, Gabriele R. Fariello, Lawrence L. Wald, Bruce Fischl, Bruce R. Rosen, Ross W. Mair, Joshua L. Roffman, Jordan W. Smoller & Randy L. Buckner

[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

Scientific Data **2**, Article number: 150031 (2015) | doi:10.1038/sdata.2015.31  
Received 29 September 2014 | Accepted 04 June 2015 | Published online 07 July 2015

[PDF](#) [ISA tab](#) [Citation](#) [Reprints](#) [Rights & permissions](#) [Article metrics](#)

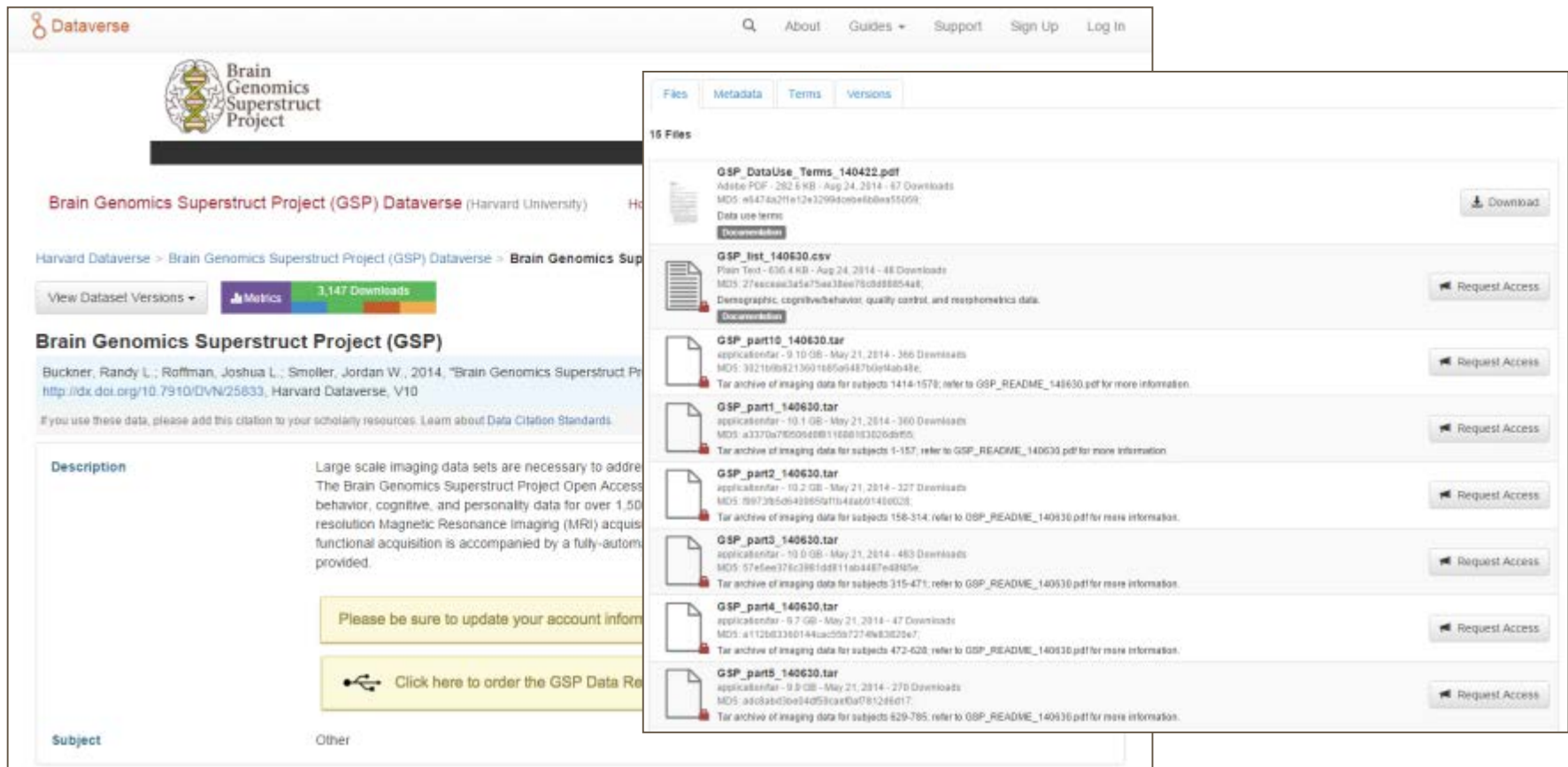
**About Scientific Data**  
Scientific Data is an open-access, peer-reviewed publication for descriptions of scientifically valuable datasets. Our primary article-type, the Data Descriptor, is designed to make your data more discoverable, interpretable and reusable.

[E-alert](#) [RSS](#)  
[Facebook](#) [Twitter](#)

**Associated Links**  
[Journal of Neurophysiology | Article](#)  
The organization of the human cerebellum estimated by intrinsic functional connectivity by A. Castellanos et al  
[Journal of Neurophysiology | Article](#)  
The organization of the human striatum estimated by intrinsic functional connectivity

<http://www.nature.com/articles/sdata201531>

# Linked to open access data



The screenshot shows the Dataverse interface for the Brain Genomics Superstruct Project (GSP). The main page displays the project title, a citation by Buckner et al. (2014), and a description of the large-scale imaging data sets. A sidebar on the right lists 16 files, including terms of use, a list of subjects, and five tar archives of imaging data for different subject groups (141-157, 158-314, 315-471, 472-620, and 629-785). Each file entry includes its format, size, date, download count, and a 'Request Access' button.

**Brain Genomics Superstruct Project (GSP) Dataverse** (Harvard University)

Harvard Dataverse > Brain Genomics Superstruct Project (GSP) Dataverse > Brain Genomics Superstruct Project (GSP)

View Dataset Versions ▾ **Metrics** 3,147 Downloads

**Brain Genomics Superstruct Project (GSP)**

Buckner, Randy L.; Roffman, Joshua L.; Smoller, Jordan W., 2014, "Brain Genomics Superstruct Project (GSP) Open Access Data", Harvard Dataverse, V10  
<http://dx.doi.org/10.7910/DVN/25833>, Harvard Dataverse, V10

If you use these data, please add this citation to your scholarly resources. [Learn about Data Citation Standards.](#)

**Description**

Large scale imaging data sets are necessary to address the question of how brain structure and function relate to behavior, cognitive, and personality data for over 1,500 subjects. The Brain Genomics Superstruct Project Open Access Data provides high resolution Magnetic Resonance Imaging (MRI) acquisition data for over 1,500 subjects. Functional acquisition is accompanied by a fully-automated cognitive and personality data collection protocol.

Please be sure to update your account information

Click here to order the GSP Data Release

**Files** Metadata Terms Versions

16 Files

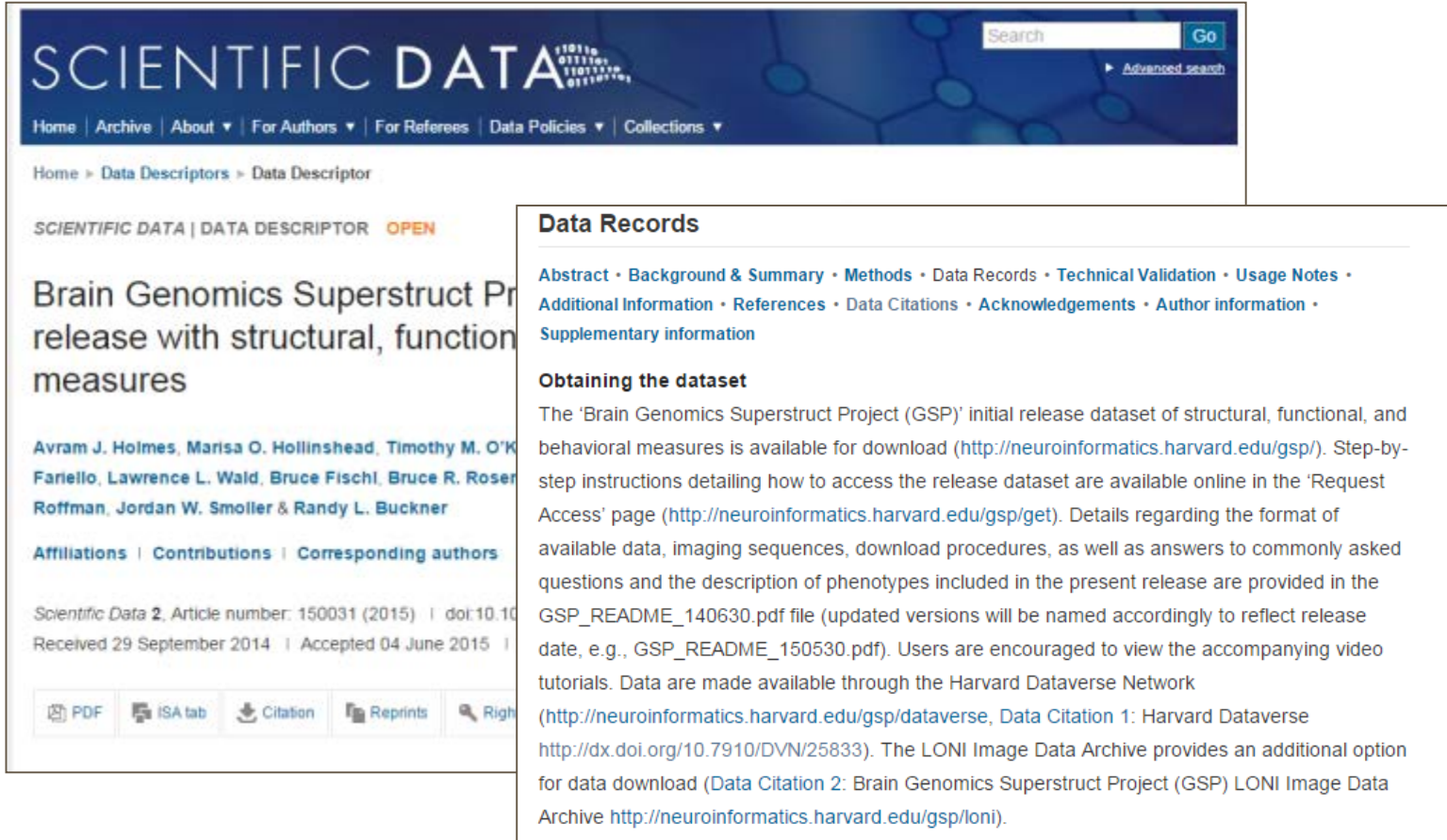
- GSP\_DataUse\_Terms\_140422.pdf**  
Adobe PDF - 262.5 KB - Aug 24, 2014 - 67 Downloads  
MD5: e6476a271e12e32994c0ebf2b0e55059  
Data use terms  
[Download](#)
- GSP\_list\_140630.csv**  
Plain Text - 636.4 KB - Aug 24, 2014 - 48 Downloads  
MD5: 274ecceae3a5e75ee38ee78c0d8885448  
Demographic, cognitive/behavior, quality control, and neurophenomics data.  
[Download](#)
- GSP\_part10\_140630.tar**  
application/tar - 9.10 GB - May 21, 2014 - 366 Downloads  
MD5: 3121b0b821360185509487b0e7ab49e  
Tar archive of imaging data for subjects 141-157; refer to GSP\_README\_140630.pdf for more information.  
[Request Access](#)
- GSP\_part11\_140630.tar**  
application/tar - 10.1 GB - May 21, 2014 - 360 Downloads  
MD5: a3370a7950948811e681b3020d9f55  
Tar archive of imaging data for subjects 1-157; refer to GSP\_README\_140630.pdf for more information.  
[Request Access](#)
- GSP\_part2\_140630.tar**  
application/tar - 10.2 GB - May 21, 2014 - 327 Downloads  
MD5: 8973b5c648865971b48a691488028  
Tar archive of imaging data for subjects 158-314; refer to GSP\_README\_140630.pdf for more information.  
[Request Access](#)
- GSP\_part3\_140630.tar**  
application/tar - 10.0 GB - May 21, 2014 - 483 Downloads  
MD5: 5745ee371c3881d8811ab4487e4845e  
Tar archive of imaging data for subjects 315-471; refer to GSP\_README\_140630.pdf for more information.  
[Request Access](#)
- GSP\_part4\_140630.tar**  
application/tar - 9.9 GB - May 21, 2014 - 47 Downloads  
MD5: a112b6336d144ac05972746b83020e7  
Tar archive of imaging data for subjects 472-620; refer to GSP\_README\_140630.pdf for more information.  
[Request Access](#)
- GSP\_part5\_140630.tar**  
application/tar - 9.9 GB - May 21, 2014 - 270 Downloads  
MD5: a0c8ab03e34d53c940a0781248e17  
Tar archive of imaging data for subjects 629-785; refer to GSP\_README\_140630.pdf for more information.  
[Request Access](#)

<http://dx.doi.org/10.7910/DVN/25833>


All approved repositories:

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

# And restricted access data



The screenshot shows the Scientific Data website interface. At the top, there is a search bar and navigation links. The main content area displays the article title "Brain Genomics Superstruct Project" and a list of authors. A callout box titled "Data Records" is overlaid on the right side of the page, providing detailed information about the dataset and how to access it.

**SCIENTIFIC DATA**  Search   [Advanced search](#)

[Home](#) | [Archive](#) | [About](#) | [For Authors](#) | [For Referees](#) | [Data Policies](#) | [Collections](#)

Home > [Data Descriptors](#) > [Data Descriptor](#)

SCIENTIFIC DATA | DATA DESCRIPTOR **OPEN**

## Brain Genomics Superstruct Project: initial release with structural, functional, and behavioral measures

[Avram J. Holmes](#), [Marisa O. Hollinshead](#), [Timothy M. O'Keefe](#), [Fariello](#), [Lawrence L. Wald](#), [Bruce Fischl](#), [Bruce R. Rosen](#), [Roffman](#), [Jordan W. Smoller](#) & [Randy L. Buckner](#)

[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

*Scientific Data* **2**, Article number: 150031 (2015) | doi:10.1038/sdata150031a  
Received 29 September 2014 | Accepted 04 June 2015

[PDF](#) [ISA tab](#) [Citation](#) [Reprints](#) [Rights](#)

### Data Records

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

#### Obtaining the dataset

The 'Brain Genomics Superstruct Project (GSP)' initial release dataset of structural, functional, and behavioral measures is available for download (<http://neuroinformatics.harvard.edu/gsp/>). Step-by-step instructions detailing how to access the release dataset are available online in the 'Request Access' page (<http://neuroinformatics.harvard.edu/gsp/get>). Details regarding the format of available data, imaging sequences, download procedures, as well as answers to commonly asked questions and the description of phenotypes included in the present release are provided in the GSP\_README\_140630.pdf file (updated versions will be named accordingly to reflect release date, e.g., GSP\_README\_150530.pdf). Users are encouraged to view the accompanying video tutorials. Data are made available through the Harvard Dataverse Network (<http://neuroinformatics.harvard.edu/gsp/dataverse>, [Data Citation 1: Harvard Dataverse](#) <http://dx.doi.org/10.7910/DVN/25833>). The LONI Image Data Archive provides an additional option for data download ([Data Citation 2: Brain Genomics Superstruct Project \(GSP\) LONI Image Data Archive](#) <http://neuroinformatics.harvard.edu/gsp/loni>).

<http://www.nature.com/articles/sdata201531>

## Appendix: Repository approval criteria

---

- Supported and recognized by scientific community
- Ensure long-term persistence and preservation of datasets
- Provide data curation
- Implement community-endorsed reporting requirements
- Provide for confidential review of datasets
- Provide stable identifiers
- Allow public access to data without unnecessary restrictions

<http://www.nature.com/sdata/about/faq#q21>

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

