# INTRO TO DATA SHARING WEBINAR *(CLICK TO LISTEN TO RECORDING)*

**HOSTED BY THE OPEN SCIENCE TASK FORCE**

**JULY 11, 2017**

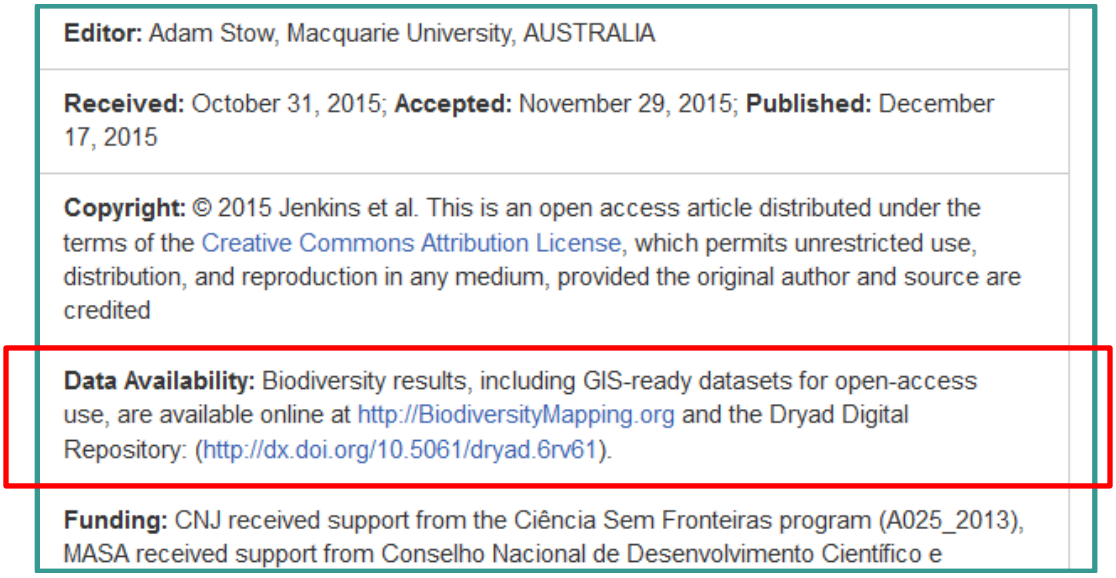
## Speaker 1: Meg Byrne, Senior Editor, PLOS ONE

## Open Data Sharing – PLOS ONE's Perspective (Slides found here.)

**PLOS Policy** (since March 2014): *(minute 6)*
www.journals.plos.org/plosone/s/data-availability
PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction (when possible).

When submitting a manuscript, authors must provide a **Data Availability Statement** describing compliance with PLOS' policy. If accepted for publication, the Data Availability Statement will be published as part of the final article.

The Data Availability Statement is openly available and machine-readable as part of the PLOS search API.



**Editor:** Adam Stow, Macquarie University, AUSTRALIA

**Received:** October 31, 2015; **Accepted:** November 29, 2015; **Published:** December 17, 2015

**Copyright:** © 2015 Jenkins et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

**Data Availability:** Biodiversity results, including GIS-ready datasets for open-access use, are available online at http://BiodiversityMapping.org and the Dryad Digital Repository: (http://dx.doi.org/10.5061/dryad.6rv61).

**Funding:** CNJ received support from the Ciência Sem Fronteiras program (A025_2013), MASA received support from Conselho Nacional de Desenvolvimento Científico e

No need to submit entire dataset, or all raw data collected, but must provide: *(minute 7)*

- the dataset used to reach the conclusions in the paper
- any additional data required to replicate the reported study findings

Minimal Dataset:

- The values behind the means, standard deviations and other measures reported
- The values used to build graphs
- The points extracted from images for analysis.

Authors are not required to make **all** images available, but sample Western Blot, fMRI image, Immunohistochemistry image, etc. must be included with the submission files or in a public repository.

Exceptions: *(minute 8)*

- Data that cannot be made publicly available for ethical or legal reasons, e.g., public availability would compromise patient confidentiality or participant privacy.
- Data deposition could present some other threat, such as revealing the locations of fossil deposits, endangered species, or farms/other animal enclosures etc.
- Data are owned by a third party.

Even when exceptions, data may be (should be when possible) available upon request to qualified researchers.

Where PLOS wants authors to put their data: *(minute 9)*

- In public repository (strongly recommended)
  Discipline-specific repositories preferable (increases findability and metadata is collected in the most appropriate manner)
  Authors must specify DOIs or accession numbers in Data Availability Statement
- Supporting information files
  PLOS can accept up to ~100 MB of data
  Each file has its own DOI and is also uploaded by PLOS to Figshare
  *(Q&A minute 49)* All PLOS journals deposit to Figshare all tables, figures, and SI files and each is given their own DOI. These are available on Figshare as well as on the PLOS platform.
- In the body of the manuscript (for small datasets)

Examples of Data Availability Statements that are (1) openly available as part of the article and (2) machine readable as part of the PLOS search API. *(minute 10)*

- EX 1 shows a DOI link that takes reader to data deposited in DRYAD.
- EX 2 shows data in supporting information which is also deposited in Figshare. Readers can get to the dataset via either Figshare or PLOS.
- EX 3 shows a restricted data example. Reasons are stated and contact info is given in the Data Availability Statement (the text of the statement is considered as part of the review process).
  - *" A data set of de-identified, population-level data is available at doi:"*
  - *"The authors will make their data available upon specific requests subject to the requestor obtaining ethical and research approvals from the Clinical Practice Research Datalink Independent Scientific Advisory Committee…"*

Outcomes: *(minute 13:30)*

Since March 2014 policy < 70,000 articles published in PLOS One with Data Availability Statements

- 60% depositing data in paper or supporting information, and deposited to Figshare
- 20% deposited directly to data repository
- 20% seeing legal or ethical restrictions on making data publically available

Initial small impact analysis *(minute 15)*

- PLOS saw an increase in data sharing from 12% to 40% of authors providing all the data needed to completely replicate the finding shown in study
- Not seeing full compliance but seeing a significant improvement
- Recent analysis saw an increase to 67% (Tim Vines, personal communication)
- The scale of PLOS journals means this is a large impact – but there is still a need to continue to increase compliance

Useful additional slides (#19 and 20) not covered but available in PLOS Blog Post (minute 16:30)

PLOS provides a wealth of additional guidance and information to authors to increase the type and amount of data authors are providing. *(minute 17)*

- Examples of PLOS ONE datasets and 3 noteworthy examples across disciplines showing the impact of sharing data and why the dataset had the impact that it did *(slide#21)*
- Data policy FAQs
- Preparing clinical data for publication
- "Ten Simple Rules for the Care and Feeding of Scientific Data"
- "Ten Simple Rules for Creating a Good Data Management Plan"
- "Sharing Research Data and Intellectual Property Law: A Primer"
- "Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data"

List of PLOS Recommended repositories *(minute 19:30)*
Based on *Nature Scientific Data (Andrew Hufton will go into more detail in next presentation)*

- Discipline specific repositories (in Biochemistry, Biomedical Sciences, Neuroscience, Structural Databases, and more)
- Cross-disciplinary repositories; Open Science Framework (OSF), Figshare, Dryad, Harvard Dataverse, and Zenodo
- Institutional repositories that adhere to best practices

PLOS continues to advocate for <OPEN DATA> *(minute 21)*

- PLOS Open Data Collection: PLOS articles selected to highlight various data practices and data practices.
    - "Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results"
    - "Sharing Detailed Research Data Is Associated with Increased Citation Rate"
    - "Ethical Challenges of Big Data in Public Health"
    - "Can Data Sharing Become the Path of Least Resistance?"
- PLOS Blog "Making Progress Toward Open Data: Reflections on Data Sharing at PLOS ONE."

Institutional support for researchers *(minute 22:30)*

- University of Cambridge provides help for preparing a research Data Management Plan
- UCSF has "DataShare" which is a data repository that UC system is creating - great for sensitive data that can't be shared publically but would benefit from being in a stable repository. UCSF Library site links to a list of NIH repositories and re3data.org which is an index of other data repositories.
-

Resources for Funders *(minute 23)*

- Implementing an Open Data Policy: A SPARC Primer for Research Funders

https://sparcopen.org/our-work/implementing-an-open-data-policy
SPARC-HRA Collaboration

**Funders can make a significant impact.** Often by the time authors submit articles to a journal, it is too late to make data shareable. If researchers are thinking about this when preparing grant applications and research proposals it can help data sharing in the long run.

Questions still remain: *(Minute 23:30)*

- How long should researchers store data?
- How much data are needed to replicate a study?
- How should materials sharing differ?
- How do we handle software/code?
- Do we need better/more aligned consenting for patient studies?
- What are best practices for data access committees?
- How can we preserve obsolete formats?
- How should data be cited and authors credited?

The good news is that many groups are thinking about these issues. Leaders in the discussion listed on Slide 27 include:
Center for Open Science home of the Open Science Framework (OSF)
ORCiD
DataCite
Nature Scientific Data
Harvard DataVerse
FORCE 11

Additional Slides were not covered in talk but they had good resources listed here:

- Data Repository Standards: Should follow FAIR Data Principles as much as possible:
  **F**indable
  **A**ccessible
  Interoperable
  **R**e-usable

- FAIRsharing.org (standards, databases, polices)

A curated, informative and educational resource on inter-related data standards, databases, and policies in the life and environmental sciences. Has recommendations, collections and education.

- re3data.org is a **R**egistry of **R**esearch data **R**epositories

Identifies over 1,500 research data repositories, making it the largest and most comprehensive registry of data repositories available on the web.

---

## Speaker 2: Andrew L. Hufton

**Managing Editor, Scientific Data (PART OF SPRINGER NATURE)**
**HRA Webinar: Intro to Data Sharing** ***(Slides found here.)***
**https://www.nature.com/sdata/**

Data sharing policy for Nature and the Nature Research journals: (*minute 28)*
http://www.nature.com/authors/policies/availability.html
Supporting data must be made available to editors and peer-reviewers at the time of submission for the purposes of evaluating the manuscript.

*"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. "*

Sharing data upon request is unreliable *(minute 28:30)*

- Vines et al found that data can be obtained only about 30% of the time, even for newest papers (even at only 2-3 years old only 40% of data could be obtained)
- Data associated with published works disappears at a rate of ~17% per year
- Datasets not referenced in a manuscript are essentially invisible (a.k.a "Dark data")
- Data producers do not get appropriate credit for their work

Over the last 5 years Nature Journals have been devoting effort to improving data sharing polices: These include but are not limited to: *(minute 20:30)*

- Including Data Availability Atatements
- Strong mandates for data sharing for specific communities
- Strong encouragements to share rich data especially in supplemental material

When developing a Data Sharing Policy: *(minute 30:30)*

- Consider how to make data **USEFUL TO OTHERS** – i.e. OPEN data is FAIR Data https://www.nature.com/articles/sdata201618
  **F**indable
  **A**ccessible
  **I**nteroperable
  **R**e-usable
- Think about **WHY** you are sharing data *(minute 31)*
  - Support data sharing within defined collaborations (i.e. sharing with friends)
  - Help others critically evaluate and reproduce an authors' claims (i.e. sharing with critics)
  - Allow others to use data in separate research projects, including overlapping or competitive research (i.e. sharing with competitors)

**The last 2 levels (critically evaluating, reproducing, and reusing data in other research) are where the real impact is.**

What is the Nature's ***"Scientific Data"***? *(minute 33 )*

- It is a peer reviewed journal where they publish data papers called "data descriptors." Data descriptors are papers that are description of datasets designed to maximize usage of data
- Citable publications that give credit for reusable data
- These "data descriptors" live between traditional results based articles and data that is on data repositories

More info at Scientific Data's webpate: https://www.nature.com/sdata/

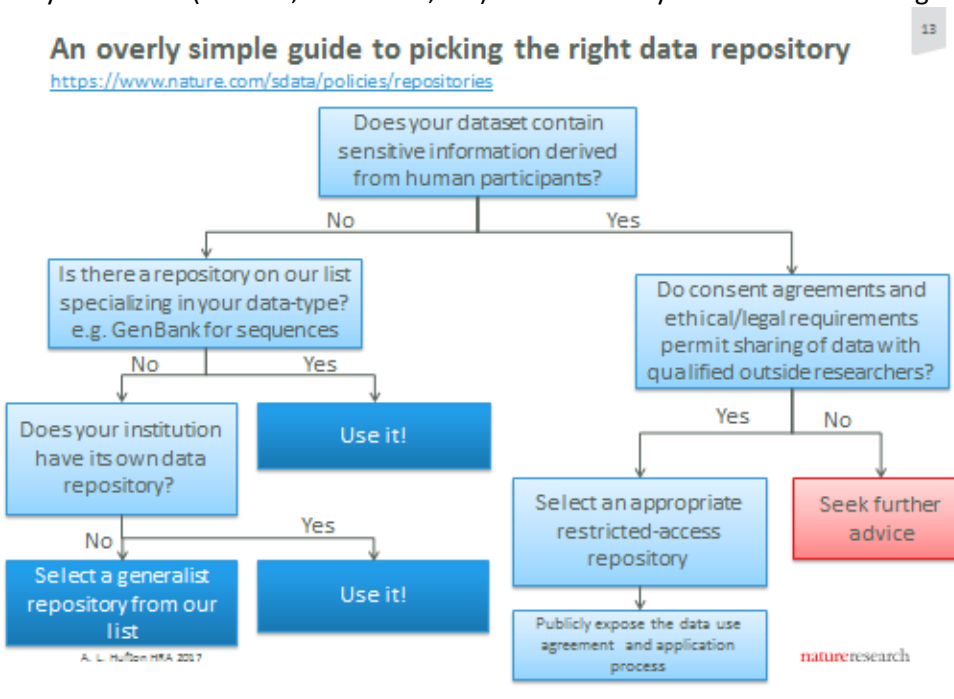
**Principles** that underlie "Scientific Data" *(minute 33:30)*

- **Get Credit for Sharing Your Data:** Publications will be indexed and citable.
- **Open-access:** Data Descriptors are published under a Creative Commons Attribution license (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:** Work with 90 different repositories. Not a new data repository; data stored in community data repositories

List of Scientific Data's recommended repositories can be found here.
What makes a good repository? *(minute 34:30)*

- Quality curation – (one that specializes in your type of data is the best)
- A commitment to long-term preservation
- Features that support collaborative analysis
- Features that allow you keep data private until you are ready to publish
- Open data licensing options (e.g. CC0 or CC BY)
  (CCO is strongly recommended for research data for legal reasons)

See this link for an overly simple guide to picking the right data repository shown below.
This is very simplified - especially on the right side. For sensitive human data - researchers should be talking to many audiences (funders, institution, etc) to ensure they follow ethical and legal guidelines.




*(minute 30:30)* **Funders can be important advisors:**
**You can ensure that grantees have a plan for how they will handle a subsequent data request.**

**Repositories that can handle sensitive human data:** *(minute 40:15)*
*(Authors should get approval/guidance from institution before depositing human data to a third-party repository. Usually, data should be carefully anonymized or pseudonymized before deposition.*

- Genotype-phenotype data archives
  - dbGAP (http://www.ncbi.nlm.nih.gov/gap)
  - EGA (http://www.ebi.ac.uk/ega/)
- National disease-specific databases
  - National Addiction & HIV Data Archive Program (http://www.icpsr.umich.edu/icpsrweb/NAHDAP/)
  - National Database for Autism Research (http://ndar.nih.gov/)
- Social science databases (they have extensive experience with human-derived datasets, and often can handle diverse kinds of clinical and health-related data)
  - UK Data Service ReShare (http://reshare.ukdataservice.ac.uk/)
  - openICPSR (http://www.openicpsr.org/)

**Generalist repositories:** *(minute 42 )*

- Figshare (www.Figshare.com)
  - In-browser data viewers, make tables and code easily previewable
  - Media files immediately playable
  - No link to a peer-reviewed publication required (can share data even before publication)
  - 100 GB of free storage available via Scientific Data, data kept private during peer-review
  - Unlimited free public storage available for researchers willing to make data immediately public
  - Great option for even large data sets - if you can make it public
  - *(Q&A Minute 50)* Figshare doesn't curate or add to info that authors provide themselves. They do provide controls for sharing specific data, searchability, and metrics for how data is being reused. BUT – Scientific Data (and many institutional repositories) DO provide curation of data that is deposited in Figshare via their repositories. Going through a public portal you get technology (DOI, controls for appropriateness and stability, etc), but going through a repository that uses Figshare you get their curation and other services.
- Dryad Digital Repository (http://datadryad.org/)
  - $120 USD for first 20 GB, and $50 USD for each additional 10 GB
  - Curation support for basic file naming & upload checking, screens for inappropriate human data
  - Fully open so this curation is key.
  - Only accepts data associated with a specific publication
- Open Science Framework (http://osf.io/)
  - Free commons that connects and integrates the entire research lifecycle, links to other platforms such as Figshare, Github, Dataverse, and many others.
- Harvard Dataverse (http://dataverse.harvard.edu/)
- Zenodo (http://zenodo.org/

**Evidence** of significant impact of publically sharing data (accidental experiment) *(minute 44)*
Original research study was published in PNAS in 2011. Data was only shared on request (no existing repository). But in 2014 data published in Figshare via Scientific Data. How much data reuse happened due to data in repository? The "Data Descriptor" has now been cited more than 94 times since 2014. This includes researchers using their data for their own studies. This data set is now used as a real benchmark for new cancer cell line screenings and bioinformatics studies.

Advice to Researchers: **Sharing data gets the most from your data.** Encourage others to do the same!

- Preserves it
- Encourages reuse
- You get credit

Another important feature of a good repository: *(Q&A minute 48)*
**Provides control:** Platforms should give researchers the ability to control with whom and when you share your data. While data is private you can generate share links and share limited data to a limited number of people, but still can push the data out and get a DOI when it goes public as part of a publication. Also the ability to set an embargo is a function of some platforms.

**This control is scientific not legal.** If you need to have legal and ethical control (especially with IP restrictions or privacy issues) you need to confer with an expert and write a robust data use agreement that is legally binding.

**How can funders incentivize data sharing?** *(Q&A minute 52:30)*

- Some funders require data sharing (Wellcome Trust and Gates)
- Communicate with researchers how data sharing can increase impact of their research
- Communicate what data sharing means to you (not just sharing within their research group)
- Require data management plans in grant proposals – but go further
- Evaluate the extent of data sharing in renewal and requests for further funding – ask them "how good of a job did you do in sharing data with people outside your research group?"
- Help cover costs. *(Q&A minute 55).* Data sharing has a very real cost to the researcher. Funders need to resource this correctly if they are going to require it. A terabyte of data can cost ~$1,000 to $20,000 to get it into a DOI-issuing repository.

**Is there a gap in repositories** for researchers in a specific disease area or a specific discipline in the landscape of repositories? Is there a need to create something that fits the nonprofit world or is something else missing? *(Q&A minute 56:30)*

Meg at PLOS noted they are seeing standards evolving constantly. There are new repositories and new ideas for data sharing added regularly. PLOS tries to be as inclusive as possible. But specific communities need to evaluate their own data needs and repositories to see if their data is discoverable and easy to reuse. If not – maybe there is a need for a new data repository.

*(Andrew - Q&A minute 58:30.)* There are significant funding differences between the US and Europe. NSF and NIH are thinking about data sharing and are putting resources in now. But they are NOT committing to long term data sharing infrastructure. There are no long-term grants on the horizon. This policy creates incentives to create many new short term data sharing solutions rather than a more robust strategy of the community coming together around a few platforms or solutions that will last.

In Europe they have pan-European solutions such as Elixir and the Open Science Cloud. They have the drawback of being very bureaucratic but are committing long-term institutional funds to durable, robust solutions.

***An important topic for funders: How can funders incentive the community to come together around a small number of resources and make sure those resources will be there for decades to come?***