Research Data Management: Sharing your Data
Hilda Teresa Ayala González

March 12, 2018
Graduate Research and Innovation Center (GRIC)
University of Puerto Rico, Mayaguez Campus

This activity is sponsored by the Transformational Initiative for Graduate Education and Research (TIGER), US Department of Education, Title V, Part B, Promoting Postbaccalaureate Opportunities for Hispanic Americans (PPOHA) Program (#P031M140035).

Pre-test

I. Why is sharing your data important

II. Mention three (3) benefits of sharing your research data

III. Mention three (3) agencies or journals that require data sharing

IV. Mention one (1) traditional way you can share your data with others

V. Mention two (2) “new” ways you can share your data with others

VI. Mention two (2) reasons why researchers (or you) would not share data
Goals

- Learn the benefits of sharing your research data
- Understand why & who requires data sharing
- Find out when & where you can share your research data
- Know the requirements for good data sharing
- Discuss the main reasons why researchers do not share data

Research Data

Data means different things to different people in different contexts. Different disciplines have and use discipline-specific language around the subject research data.

- Data may be viewed as the lowest level of abstraction from which information and knowledge are derived.
Research Data Lifecycle

The data lifecycle provides a high level overview of the stages involved in successful management and preservation of data for use and reuse.

University of Virginia, n.d.; DataOne, n.d

Research Data

Data are the building blocks of empirical research, whether in the behavioral, social, biological, or physical sciences. To understand fully and extend the work of others, researchers often require access to the data on which that work is based. Yet many members of the scientific community are reluctant or unwilling to share their data even after publication of analyses of them. Sometimes this unwillingness results from the conditions under which data were gathered; sometimes it results from a desire to carry out further analyses before others do; and sometimes it results from the anticipated costs, in time or money, or both.

The Committee on National Statistics believes that sharing scientific data with colleagues reinforces the practice of open scientific inquiry. Cognizant of the often substantial costs to the original investigator for sharing data, the committee seeks to foster attitudes and practices within the scientific community that encourage researchers to share data with others as much as feasible.

1985 National Academy of Sciences

National Research Council, 1985
WHO, HOW, WHAT, WHEN, WHERE, WHY

WHO is involved: Stakeholders

- Institution
- Creators/Researchers
- Sponsors
- Users
- Publishers
Principal Stakeholders

Sponsors

Publishers

Sponsors
Sponsors: Data Availability

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis, and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

February 22, 2013

GEND, n.d.
Sponsors: Data Management Plan

A data management plan is a formal document that outlines what you will do with your data during and after a research project. Most researchers collect data with some form of plan in mind, but it’s often inadequately documented and incompletely thought out.

Checklist for a Data Management Plan
Digital Curation Center

Data Management = Good Research --> Sharing
Publishers

Publishers Findings

• Provenance
• Replication
• Transparency
• Reproducibility
• Access

POSSIBLE RED FLAGS OF RESEARCH MISCONDUCT

TIME
- Data are too good to be true
- Findings can't be replicated by others in the lab

RESULTS
- Usable data are only generated when there is a pricing deadline
- Experiments are completed faster than usual

LACK OF TRANSPARENCY
- Raw data can't be produced when requested
- Research materials and protocols are kept hidden
- Work is mostly done when no one else is around

If you suspect research misconduct, contact your institution's Research Integrity Officer or ORI at ORIedu

Office of Research Integrity, n.d. Possible
Publishers: Reproducibility

Nature 2016 study found “More than 70% of researchers have tried and failed to reproduce another scientist’s experiments, and more than half have failed to reproduce their own experiments”.

Publishers: Reproducibility

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?
Most scientists have experienced failure to reproduce results.

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.

Baker, May 2016
Publishers: Sharing Research Data

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.

WHY Should Research Data be Shared

- Encourages scientific enquiry and debate
- Promotes innovation and potential new data uses
- Leads to new collaborations between data users and data creators
- Maximizes transparency and accountability
- Enables scrutiny of research findings
- Encourages the improvement and validation of research methods

- Reduces the cost of duplicating data collection
- Increases the impact and visibility of research
- Promotes the research that created the data and its outcomes
- Can provide a direct credit to the researcher as a research output in its own right
- Provides important resources for education and training
WHAT Data Should be Shared

- Requirements from sponsors or journals
  - Read policies before applying/submitting
  - No privacy, PII, sensitive or commercial/secrecy issues

*Also include documentation about your data!*

- Research data needed to validate findings
  - For public scrutiny

- Valuable data for others
  - Current & future

- Unique data
  - Cannot be re-generated

WHEN Should your Data be Shared

- As established by Sponsors or Journals
  - Read policies
  - No privacy, PII, sensitive or commercial/secrecy issues

**NSF DATA SHARING POLICY**

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See Award & Administration Guide (AAG) Chapter VI.D.4.

**Timeframe for Data Sharing**

Data sharing should be timely and no later than the acceptance for publication of the main findings from the final dataset. Data from large studies can be released in waves as data become available or as they are published.

**Springer Nature**

A condition of publication in a Nature journal is that authors are required to make unique materials promptly available to others without undue qualifications.
**WHEN** Should your Data be Shared

Restrictions:
- Patent (IP) application
- Confidential data
- Sensitive data
- National security

You can:
- Embargo
- Informed consent
- Anonymized
- Limit access by request

---

**Where & how do you share your data?**
WHERE Should your Data be Shared

**Classical way**
- Thesis, Dissertation or Article
  - Research results
- Poster or Conference presentation
  - Research results

**Ok way**
- A link to data set
- As supplemental material
  - Not such a good idea

Unstructured repositories like [figshare](http://figshare.com) and [Dryad](http://dryad.org) are suitable alternatives if no structured public repositories exist. As a less desirable alternative, data sets can be made available as Supplementary Information files, which will be freely accessible on nature.com upon publication. In rare cases when data files cannot be deposited in an accessible repository for technical reasons, authors must make the data available to editors and peer reviewers if requested. After publication, authors must likewise arrange to make the data available to any reader directly upon reasonable request.

---

**Encouraged way**
- Data Repository
  - Discipline specific
  - Institutional/Public
  - General purpose

[GenBank](http://www.ncbi.nlm.nih.gov/Genbank/)
[The Dataverse Project](http://www.dataverse.org/)
[DRYAD](http://www.dryad.org/)
[figshare](http://figshare.com/)
[Center for Open Science](http://cos.io/)
[PANGEA](http://pangea.is/)
[DataONE](http://www.dataone.org/)
[ICPSR](http://www.icpsr.umich.edu/)
[QDR](http://www.qdrdata.org/)
[Interuniversity Consortium for Political and Social Research (ICPSR)](http://www.icpsr.umich.edu/)
WHERE Should your Data be Shared

Data Repository
- Trained personnel!!
- Provide a time range for keeping your data
- Metadata = Discoverability
- Persistent identifiers = Retrievable
- Migration of file formats = Reproducibility
- Licensing = Standard policies for re-use
- Citation tools = Ensures credit over data
- Metrics = Knowing the impact of your data

Storing IS NOT Preserving
WHERE Should your Data be Shared

New trend: Data Journals or Data Articles

- Published details about data sets:
  - Nature’s Scientific Data
  - Bio-Medical Central
  - Dataset Papers in Science
  - Journal of Open Archaeology Data
  - Open Data Journal for Agricultural Research
  - Elsevier’s Data in Brief
  - See more

Higher visibility + Increase usage --> Data Re-use

HOW Should your Data be Shared

Plan

- Know your grounds
  - Mandates, policies, requirements

- Find a repository
  - Adequate to your data & needs
  - Requirements
    - Formats
    - Metadata
    - Supplemental documentation
    - Cost

- Manage your data
  - Type of data, formats, naming scheme, responsibilities, software, storage, security, access & use policies

- Think of ownership & privacy
  - Sensitive data, PI, copyrights

- Establish a sharing schedule
  - When the data will be available to others
HOW Should your Data be Shared

Prepare

- Clean it
  - Make sure variables are well identified
  - Eliminate unnecessary details
  - If sensitive or PII information
    - Anonymize it: remove direct identifiers, reduce precision of information, use pseudonyms
- Use non proprietary formats
  - csv vs excel
- Assign a license
  - The less restrictive the better
    - CC0 / PDDL
- Follow the 5 stars to open data!

your data

How Should your Data be Shared

Share it!

- Deposit your data
  - Write an article about it too!
- Get an identifier
  - Digital Object Identifier (DOI)
    - Is an alphanumeric string assigned to uniquely identify an object
- Citation
  - Get attribution & credit for your effort
    - Creator (PublicationYear). Title. Version. ResourceType. Publisher. Identifier
  - DOI Citation [Formatter]
**HOW to Find Shared Data**

**Exercise**

- Go to [http://www.re3data.org/](http://www.re3data.org/)
- Browse -> Browse by Subject
- Select your discipline
- Identify a repository

**What are the reasons that might inhibit you to share your data?**
Wiley’s Researcher Data Sharing Insights Survey (2014)

If the investigators believed that these advocates were interested in advancing scientific knowledge, the investigators would be willing to share, but believing that the advocates are only interested in looking for weaknesses in the data that can be used to discredit the researchers makes them hesitant to make the data public.  
Meadows, 2014

Research Data

Progress in science and engineering depends on the collection of data through observation, experimentation, and computation. A core expectation of the scientific method is the documentation and sharing of results, underlying data, and methodologies. This process enables other researchers to reproduce experiments and studies, verify and validate results, and build upon previous work to produce further scientific advances.

National Science Foundation, 2011
Sharing Research Data
Is good for you too!

Questions and comments
Post-test

I. Why is sharing your data important

II. Mention three (3) benefits of sharing your research data

III. Mention three (3) agencies or journals that require data sharing

IV. Mention one (1) traditional way you can share your data with others

V. Mention two (2) “new” ways you can share your data with others

VI. Mention two (2) reasons why researchers (or you) would not share data

References


Data Cite. (n.d.). Why is it so important to cite data? Retrieved from https://www.datacite.org/cite-your-data.html


References

University of Oxford. (n.d.). To share or not to share? Retrieved http://researchdata.ox.ac.uk/home/sharing-your-data/to-share-or-not-to-share

Evaluation: http://uprm.libsurveys.com/tiger
Password: 2016

Research Data Management: Sharing your Data
Hilda Teresa Ayala González
March 12, 2018
Find Us

http://libguides.uprm.edu/gric

gric@uprm.edu

@GRICUPRm

@GricUprm

GRIC UPRM