



Sharing and Reusing Research Data

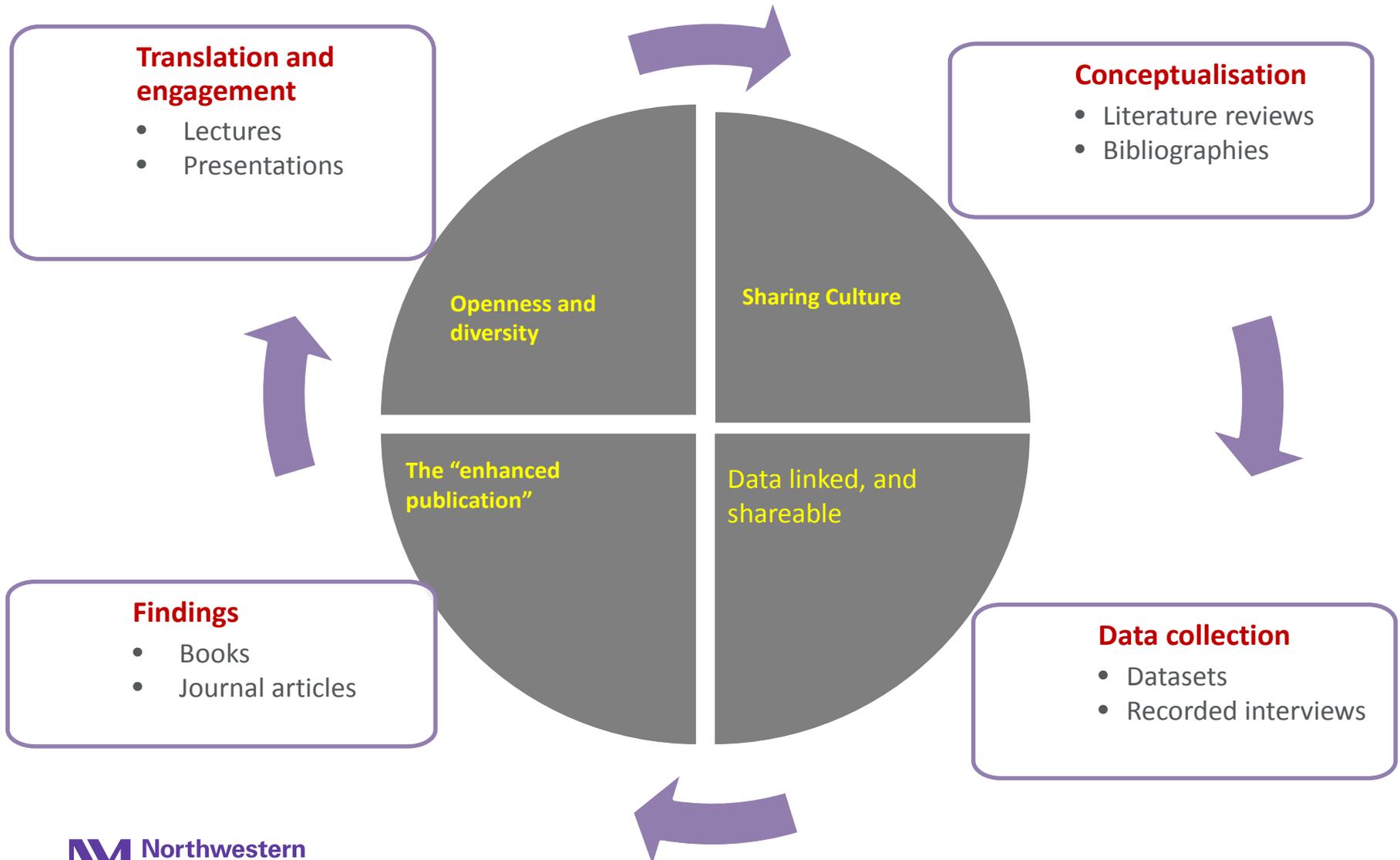
Ehsan Mohammadi, PhD

Postdoc Fellow, Preventative Medicine-Health and Biomedical Informatics
Northwestern University, Feinberg School of Medicine

Overview

- Scholarly communication cycle
- Why share data?
- Advantages of data sharing
- Ethical issues for data sharing
- Barriers for data sharing
- Where share / search data?
- Summary

The changing scholarly communication cycle



Scholarly communication is changing



Why share data?



- This is my data....., my assets
- This is my data...., secure my team's future careers
- This is my data..., make me a strong candidate for future funding, why should I endanger this?
- This is my data..., other researchers can not do better with my data?

Definition of data sharing?

- Sharing data is not a new topic.
- “Data sharing is the release of research data for use by others” (Borgman, 2012)
- Starts from the 1980s (Fienberg, Martin, & Straf, 1985)
- Requirements to deposit genome sequences are well recognized. (Wellcome Trust, 1996)

The life of data

Data can be used and re-used for future research, if:

- ✓ shared
- ✓ managed well
- ✓ properly preserved
- ✓ made available

http://eprints.ncrm.ac.uk/2846/1/DMPlanning_Presentation.ppt

Advantages of data sharing?

- Reduces duplication of effort
- Increases the impact of research (70% more)
- Provides credit to data as a research contribution
- The success of research is measured by the number of citations it receives
- Provides a means of sharing research data



OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

Sharing Detailed Research Data Is Associated with Increased

308

Comments	Related Content
----------	-----------------

Citation analysis: A study of seldom-cited influences
MacRoberts
2009 Full publication history
/view/save citation
Citation tools

View issue TOC
Volume 61, Issue 1
January 2010
Pages 1–12

i cancer microarray clinical trial publications with respect % of trials with publicly available microarray data received by available data was significantly (p=0.006) associated with the tendency of journal impact factor, date of publication, and progression.

able data and increased literature impact may further detailed research data.

Outputs of the NISO Alternative Assessment Metrics Project

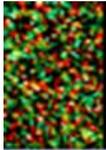
ences on the production of a scientific article, the content of the article
/e examined articles in biogeography and found that most of the
ed, specific types of articles that are influential are cited while other
re influential are not cited, and work that is “uncited” and “seldom cited”
. As a result, evaluative citation analysis should take uncited work into

Advantages of data sharing?

- Outputs of the funded research should maximize public benefit.
- Make research data timely and widely available to the research community
- Share in responsible manner ensuring data can be verified, built upon, and used to advance knowledge

http://grants.nih.gov/grants/regionalseminars/2014/docs/Presentations2014/Inventions_Data%20Sharing_and_other_%20IP_Considerations.pptx

Why share data?



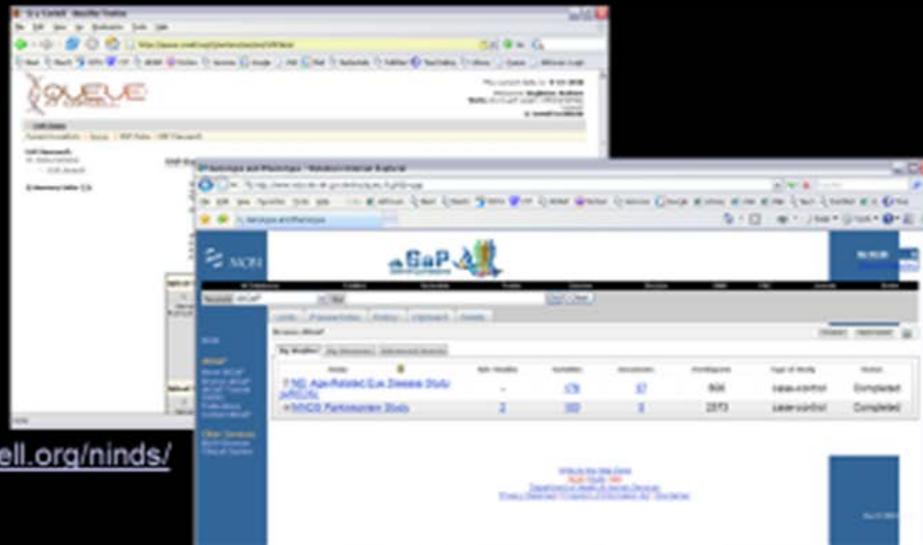
Andrew Singleton
Molecular Genetics Section
Laboratory of Neurogenetics
National Institute on Aging
singleton@mail.nih.gov



Public release – facilitating discovery

>600,000,000
genotypes from PD,
ALS stroke and control
cohorts posted publicly

downloaded by >500
unique visitors



<http://ccr.coriell.org/ninds/>

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap>

Two manuscripts published; five in
press/preparation by 'downloaders'

Audiences of shared data

- Other academics
- Government
- NGOs
- Private companies
- Whole community.

Why is data sharing important?

Is data sharing a buzz word?



What is your idea?

<input checked="" type="checkbox"/>	YES
<input type="checkbox"/>	NO
<input type="checkbox"/>	MAYBE

A close-up photograph of a hand holding a red pen, pointing at the 'YES' option in a survey form. The form has three rows, each with a square checkbox and a corresponding label: 'YES', 'NO', and 'MAYBE'. The top checkbox is checked with a red mark, and the red pen is positioned just above the 'YES' label.



Why share data?



Added a data management plan requirement in March, 2003 for grants over \$500K



Required a data management plan for all full proposals submitted January, 2011

And other organizations around the world



Why share data?

- Deposit of data and other research documentation associated with published articles are encouraged.
- Journals in economics and many other fields also require access to data.
- *Science* encouraged scholars to share data

The screenshot displays the Science journal website interface. At the top, the 'Science' logo is prominent, with 'AAAS' to its right. Navigation links for 'Home', 'News', 'Journals', 'Topics', and 'Careers' are visible in a red bar. Below this, a search bar and a list of journal categories are shown. The main content area features an article titled 'Making Data Maximally Available' by Brooks Hanson, Andrew Sugden, and Bruce Alberts. The article is categorized as an 'EDITORIAL'. Social media sharing options for Facebook, Twitter, and Email are provided. A 'SHARE' button is also present. Below the article title, there are tabs for 'Article', 'Figures & Data', 'Info & Metrics', 'eLetters', and 'PDF'. A red button labeled 'View Full Text' is located at the bottom right of the article summary. To the right of the article, there is a sidebar with a 'Science' logo and a 'data' graphic, along with a list of 'ARTICLE TOOLS' including 'Download Powerpoint', 'Save to my folders', 'Request Permissions', and 'Share'. The text 'You are currently viewing the summary.' is displayed at the bottom left of the article area.



Why is data sharing important?

- Required by publishers
- Data as a public investment
- Required by government funding agencies
- Informs new research
- Maximizes transparency, accountability and scrutiny of research findings

<https://goo.gl/scOez8>

Ethical issues of data sharing



Research ethics in data creation

The researcher has

- Duty of confidentiality towards participants
- Duty to protect participants from harm
- Duty to treat participants as intelligent, capable and able to make their own decisions
- Duty to explain the aims of the research
- Duty to inform participants how information and data obtained will be used in the short-term

<https://goo.gl/tb84wU>

Research ethics in data use

The Researcher has;

- duty to inform participants how data obtained will be used in the long-term
 - *Processed*
 - *Added to a database*
 - *More widely shared*
 - *Disposed of, etc.*
- Duty of use to accurately reflect the information gathered
- Duty to wider society to make available resources produced with public funds (e.g. data sharing required by research councils).

<https://goo.gl/tb84wU>



Access restrictions

- The two key reasons for access restrictions are:
 - to protect confidentiality of participants in the data collection and
 - intellectual property concerns.



Data sharing barriers

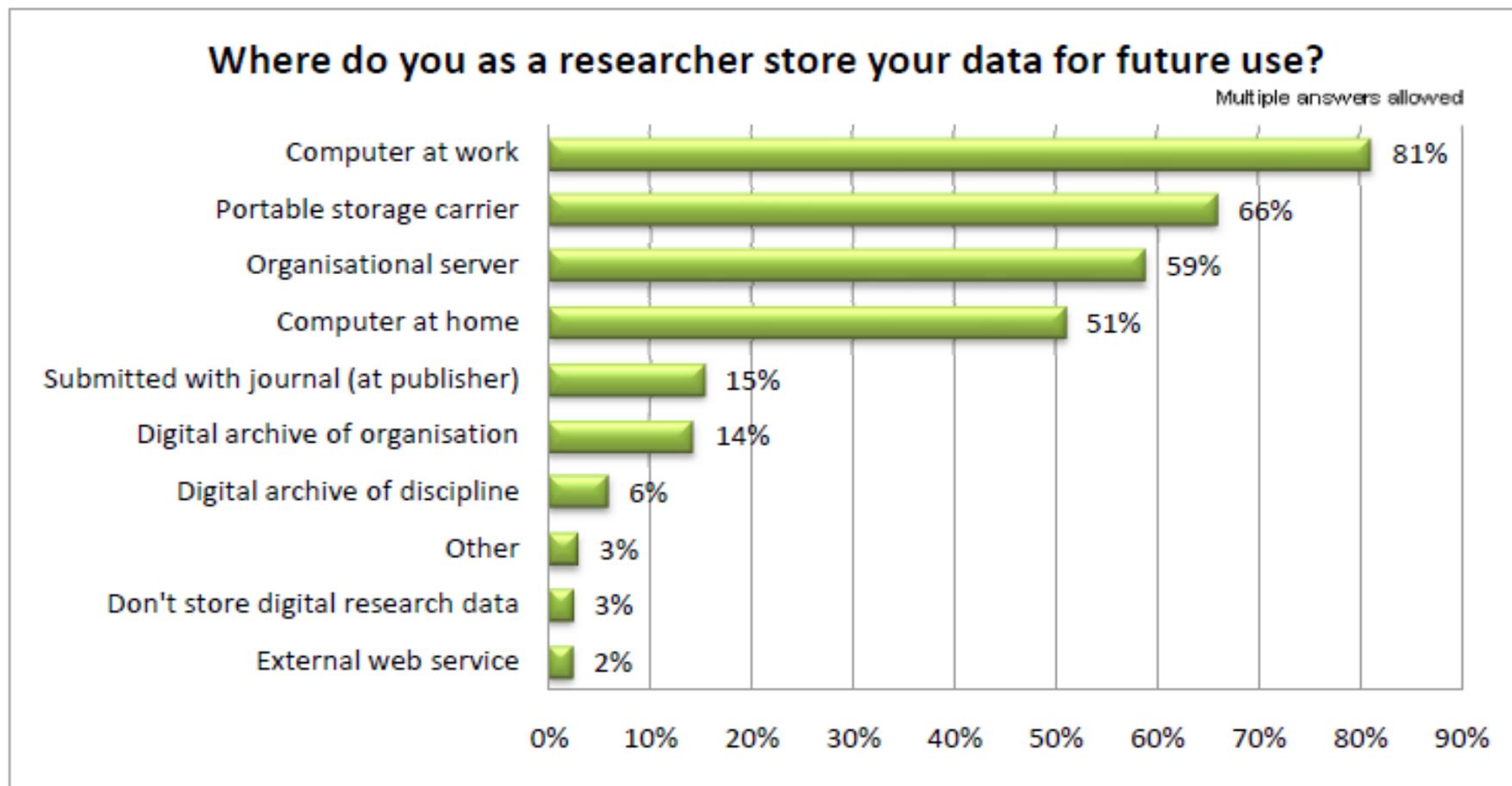
- Priority of published papers
- Little academic reward for development and sharing of datasets/DB
- Required effort to share re-usable data (incl. formatting, metadata creation, licensing etc.)
- Existing copyrights, confidential and sensitive data
- Concerns that data could be scooped, misused or misinterpreted
- Potential reputational risk (e.g. data quality, errors,...)

http://www.ariadne-infrastructure.eu/fre/content/download/8075/47364/file/ARIADNE_EAA-Vilnius_OpenDataRequirements_GGeser_01092016.ppt

Tools for depositing and searching data



Where is research data stored?

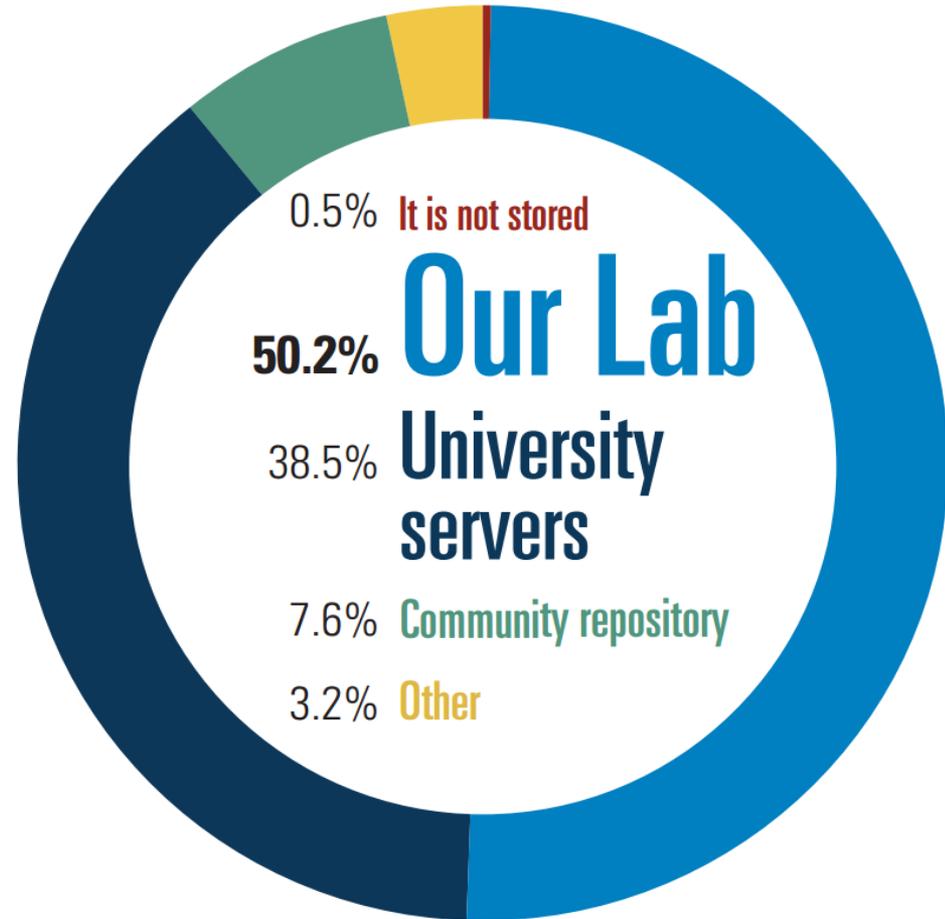


<https://goo.gl/057Nv7>

Where is research data stored?

Where do you archive most of the data generated in your lab or for your research?

“ Even within a single institution **there are no standards for storing data**, so each lab, or often each fellow, uses ad hoc approaches. ”



<http://science.sciencemag.org/content/331/6018/692>



What Should My Deposit Include?

1. **Data File:** SAS, SPSS, or Stata files.

- ✓ Each variable in the data collection should have a set of exhaustive, mutually-exclusive codes. Variable labels and value labels should clearly describe the information or question recorded in that variable. Missing data codes should be defined.
- ✓ Remove identifying information from the data to protect confidentiality.

<https://www.icpsr.umich.edu/icpsrweb/content/civicleads/deposit.html>



What Should My Deposit Include?

2. Documentation

- ✓ Documentation should thoroughly explain the data collection. Examples of documentation files include codebooks, data collection instruments, summary statistics, project summaries, and bibliographies of publications pertaining to the data.

3. Description

- ✓ Study descriptions are valuable resources to data users, and include both general information, such as study title and summary, as well as more detailed information about the study design and methodology.

<https://www.icpsr.umich.edu/icpsrweb/content/civicleads/deposit.html>

Where should I deposit my data? NCBI

NCBI Resources How To ehsan@northwestern.edu My NCBI Sign Out

NCBI National Center for Biotechnology Information

All Databases Search

NCBI Home

Resource List (A-Z)

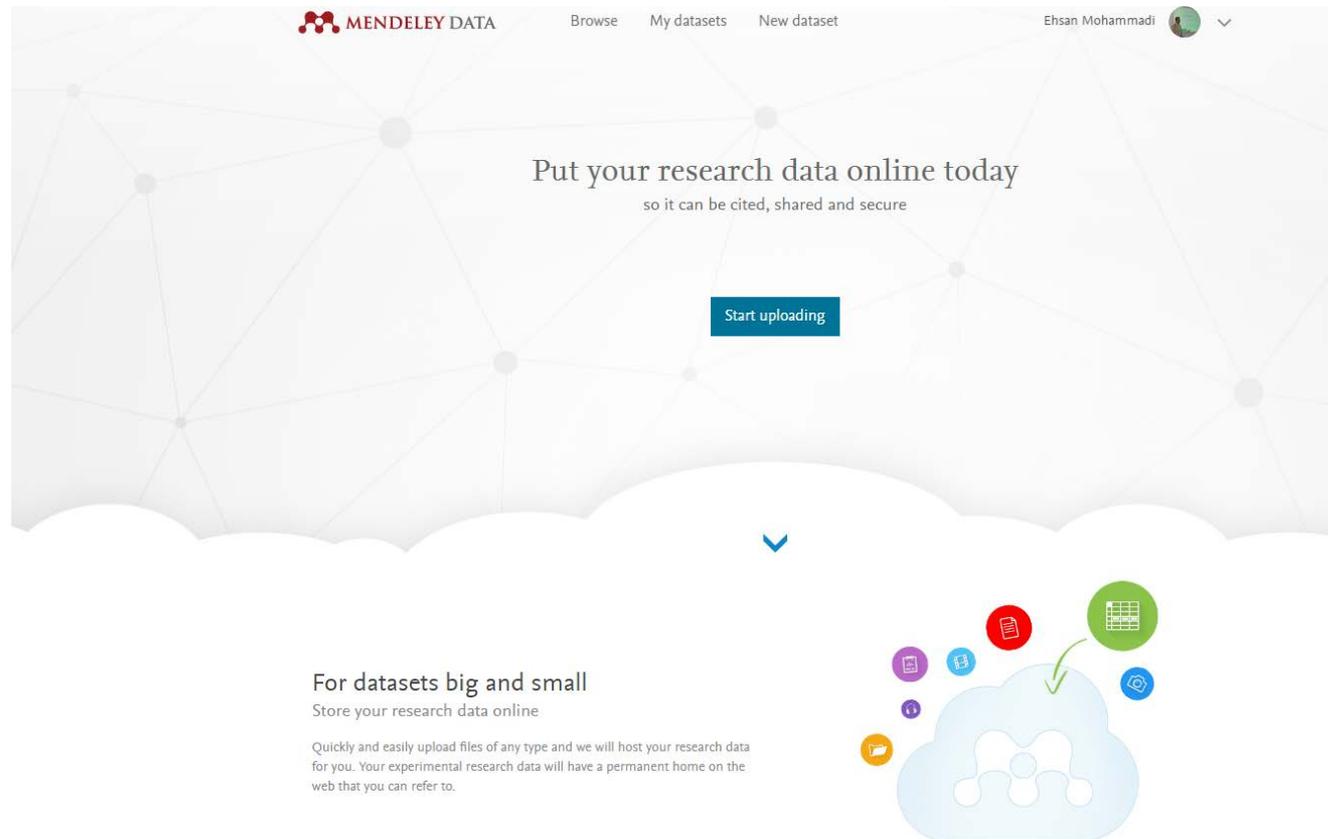
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

How to: Submit sequence data to NCBI

Starting with...	NOTES	SUBMISSION TOOLS & HELP DOCUMENTS
Simple Sequence Submissions		
Single nucleotide sequence or Several nucleotide sequences for <i>different</i> genes or loci	Contiguous bases of cDNA or genomic DNA, but should not be complete genomes. Complete genomes should be submitted via the appropriate protocol indicated below. Records with simple annotation may be submitted by BankIt or Sequin , while records with complicated annotation may be more easily submitted via Sequin .	BankIt or Sequin
Group of nucleotide sequences for the <i>same</i> gene or locus	Includes: <ul style="list-style-type: none"> population studies (sequences for a single organism) phylogenetic studies (sequences for multiple organisms) environmental samples (such as cultured or uncultured bacteria or metagenomic samples) 	BankIt or Sequin
	Includes:	Batch submit evidence

<https://www.ncbi.nlm.nih.gov/guide/howto/submit-sequence-data/>

Where should I deposit my data?



MENDELEY DATA

Browse My datasets New dataset

Ehsan Mohammadi

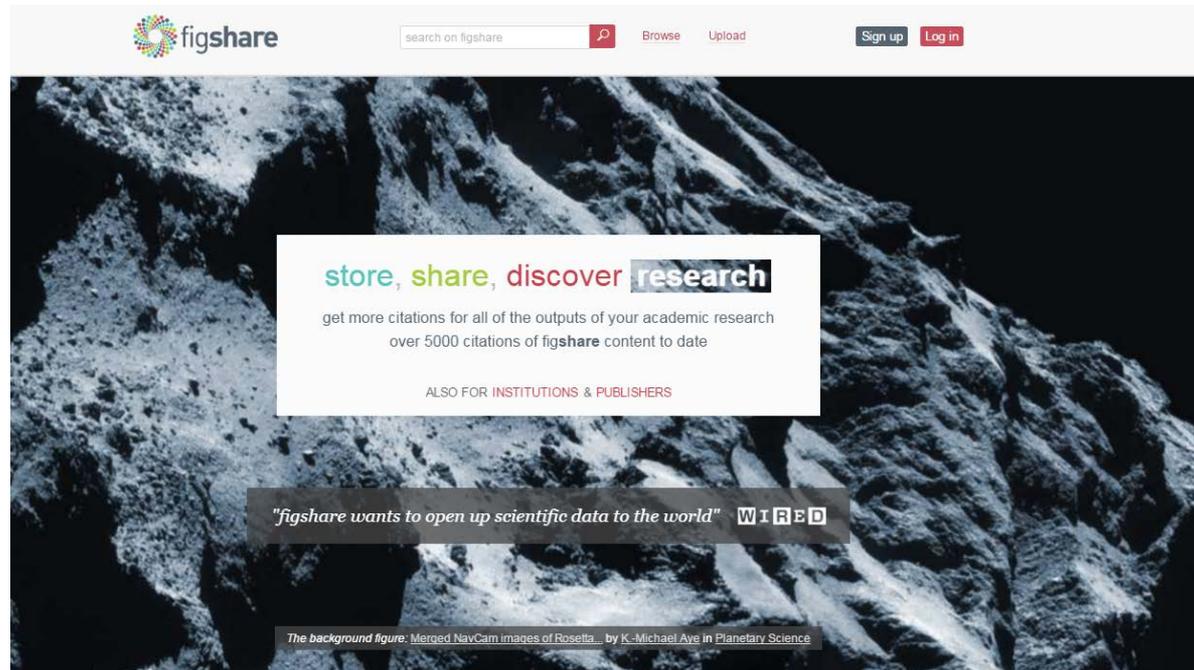
Put your research data online today
so it can be cited, shared and secure

Start uploading

For datasets big and small
Store your research data online

Quickly and easily upload files of any type and we will host your research data for you. Your experimental research data will have a permanent home on the web that you can refer to.

Where should I deposit my data?



simplifying your research workflow

easily manage all your research outputs and make them available in a citable, shareable and discoverable manner

<https://figshare.com>

Where should I deposit my data? (DigitalHub)

Step1: Login digitalhub.northwestern.edu

Northwestern Medicine

Northwestern University | Faculty Profiles

Northwestern University Feinberg School of Medicine

DigitalHub BETA

Home Catalog About News Help Contact

Enter search terms All

Log in

NetID

Password

Remember me

Log in

Where should I deposit my data? (DigitalHub)

Step2: Upload

Signed in successfully. ✕

My Dashboard

Hello, Violeta Ilik



Upload



Create Collection



View Collections



View Files

Your Statistics

User Activity

digitalhub.northwestern.edu/dashboard

Where should I deposit my data? (DigitalHub)

Step3: deposit agreement

Upload

My Computer

Cloud Providers

Note:

You must agree to [DigitalHub's Deposit Agreement](#) before starting your upload.
You must have the rights to let Galter Library distribute it.

On Behalf of Yourself

Upload to: Select collection...

I have read and do agree to [DigitalHub's Deposit Agreement](#).

+ Select files...

+ Select folder...

Start upload

Cancel upload

Where should I deposit my data? (DigitalHub)

Step4: Start Upload

Upload

 My Computer

 Cloud Providers

Note:

You must agree to [DigitalHub's Deposit Agreement](#) before starting your upload.
You must have the rights to let Galter Library distribute it.

On Behalf of

Upload to:

I have read and do agree to [DigitalHub's Deposit Agreement](#).

+ Select files...

+ Select folder...

 Start upload

 Cancel upload

Where should I deposit my data? (DigitalHub)

Step5: Apply Metadata

5 required fields:

Title

Resource type

Keyword

Creator

Rights

*** Title ?**

DigitalHub: Preserving and Sharing Your Work Using NM's – Remove

+ Add

*** Resource type(s) ?**

Poster
Practice Guideline
Pragmatic Clinical Trial
Presentation
Price Lists
Problems and Exercises
Programmed Instruction
Programs
Project

*** Keyword ?**

Institutional Repository – Remove

+ Add

*** Rights ?**

Attribution-ShareAlike 3.0 United States – Remove

Click to select... + Add

*** Creator ?**

Ilik, Violeta – Remove

+ Add

Where should I deposit my data? (DigitalHub)

Step6: visibility & permissions

Bulk Permissions (applied to all files just uploaded)

Permissions are **not** saved until the "Save" button is pressed at the bottom of the page.

Visibility - who should have the ability to read and download ?

- Open Access (recommended)** Visible to the world. **Northwestern University** Visible to all Northwestern University users.
 Private Visible to users/groups specified below, if any.

Share With (optional) ?

Enter User (one at a time)

Holmes, Kristi Lee (klh183) ▼

Galter Health Science Library; Feinberg School of Medicine ▼

Choose Access ▼



Choose Access ▼



Person/Group	Access Level
Depositor (Violeta Ilik)	Edit
klh183	Edit

Where should I deposit my data? (DigitalHub)

Step7: Save!

Your files are being processed by DigitalHub in the background. The metadata and access controls you specified are being applied. Files will be marked **Private** until this process is complete (shouldn't take too long, hang in there!). You may need to refresh your dashboard to see these updates.

Upload

Create Collection

My Files

My Collections

My Highlights

Files Shared with Me

Sort By: relevance Show 10 per page

Refresh

Filter your files

Congratulations, you have submitted your work!

How to find data for future studies?

Subject repositories

See here for a list across different disciplines

http://oad.simmons.edu/oadwiki/Data_repositories

How to find data for future studies?



<https://www.datacite.org/>

Search for work

Search

7,370,444 Works**Comportamiento de conductores eléctricos usados en líneas de transmisión ante esfuerzos electromecánicos y térmicos combinados**

Sandra Milena Téllez Gutiérrez

Work published 2011 via Unpublished

<https://doi.org/10.13140/RG.2.2.24400.71684> Cite**New draft item**

Sebastian Hormigo

Fileset published 2016 via Figshare

This is the metadata used for the analysis presented in the manuscript entitled Tuned-up%0Asomatosensory filtering neurological processes in athletes relate to better%0Aoutcomes in physical conditioning tests. By Sebastian Hormigo, Antonio Cardoso, Consuelo Sancho, Dolores E. López,%0Aand Carlos Moreno.

<https://doi.org/10.6084/Mg.FIGSHARE.4036008.V2> Cite**Disclosure Practices of Mobile Telecommunication Companies with Special Reference to Grameenphone Ltd.**

Taposh Kumar Neogy (Institute of Business Administration (Under National University), Rajshahi, BANGLADESH)

Dataset published 2016 via Harvard Dataverse

<https://doi.org/10.7910/DVN/IQKAKM> Cite**GBIF Occurrence Download**

occdownload.gbif.org

Dataset published 2016 via The Global Biodiversity Information Facility

A dataset containing 16 species occurrences available in GBIF matching the query: TaxonKey: Mascagnia schunkei W.R. Anderson. The dataset includes 16 records from 7 constituent datasets: 2 records from Naturalis Biodiversity Center (NL) - Botany, 1 records from HAMAB - Herbário Amapaense, 3 records from Herbarium - Instituto Nacional de Pesquisas da Amazônia (INPA), 3 records from Tropicos Specimen Data, 2 records from NMNH Extant Specimen and Observation Records, 4 records from Field Museum of...

<https://doi.org/10.15468/DL.IJZ21M> Cite**Resource Types**

<input type="checkbox"/> Dataset	2,835,764
<input type="checkbox"/> Text	1,437,310
<input type="checkbox"/> Other	878,712
<input type="checkbox"/> Image	713,474
<input type="checkbox"/> Collection	360,419
<input type="checkbox"/> Software	17,452
<input type="checkbox"/> Audiovisual	8,413
<input type="checkbox"/> Event	6,764
<input type="checkbox"/> Physical object	6,680
<input type="checkbox"/> Film	935
<input type="checkbox"/> Model	599
<input type="checkbox"/> Interactive resource	398
<input type="checkbox"/> Sound	251
<input type="checkbox"/> Workflow	238
<input type="checkbox"/> Service	25

Publication Year

<input type="checkbox"/> 2016	653,033
<input type="checkbox"/> 2015	2,045,544
<input type="checkbox"/> 2014	962,919
<input type="checkbox"/> 2013	341,085
<input type="checkbox"/> 2012	220,608
<input type="checkbox"/> 2011	343,603
<input type="checkbox"/> 2010	150,422
<input type="checkbox"/> 2007	164,179
<input type="checkbox"/> 2006	149,947
<input type="checkbox"/> 2005	165,340

Data Centers

<input type="checkbox"/> Data-Planet	865,742
<input type="checkbox"/> The Cambridge Crystallographic Data Centre	623,107
<input type="checkbox"/> E-Periodica	513,621
<input type="checkbox"/> TU Lododsmuseum	487,448
<input type="checkbox"/> figshare ARS	444,342
<input type="checkbox"/> ResearchGate	400,824
<input type="checkbox"/> Deutsches Institut	373,193

Basic tips for a data management plan – start with these questions

- What data will be shared?
- Who will have access to the data? (public/private?)
- Where will the data to be located?
- When will the data be shared?
- What format the data should be stored?
- How will researchers locate and access the data?
- What is data retention/destruction policy?
- What control do you have over data once it leaves your department?
- What kind of follow-up is appropriate?
- How can you be more transparent about how your data is used?
- Do you have a guidelines for version control and file naming?

Getting started...

The [Data Management Planning \(DMP\) Tool](#)

developed by the California Digital Library provides step-by-step instructions.

There are many good examples of data management plans on university websites like [University of Michigan](#), [University of Minnesota](#), and [NCSU](#)

Summary

- Data sharing is encouraged at national and global levels.
- Data sharing is becoming more standard.
- Data sharing would be easier than before with developed new technical infrastructures.
- Data sharing repays the investment to the society.
- Data sharing helps knowledge advancement.
- Data sharing increases research impacts and scientific collaborations.

References and Further Resources

- Berkeley's Research Data Management Program
<http://researchdata.berkeley.edu/about>
- Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059–1078.
- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Mit Press.
- Czerniewicz, L., Kell, C., Willmers, M., & King, T. (2014). Changing Research Communication Practices and Open Scholarship: A Framework for Analysis.
- Data Management Planning Tool <https://dmptool.org/>
- Data protection and research data <https://www.jisc.ac.uk/guides/data-protection-and-research-data>
- Data sharing with Molecular Genetics focus:
<https://www.youtube.com/watch?v=Bh6FQjYONEY>

References and Further Resources

- European Commission: Online survey on scientific information in the digital age, Brussels, 2012, http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf
- European High-level Expert Group on Scientific Data (2010): Riding the wave. How Europe can gain from the rising tide of scientific data. A submission to the European Commission, October 2010, <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>
- Fienberg, S.E., Martin, M.E., & Straf, M.L. (Eds.). (1985). Sharing research data. Washington, DC: National Academies Press
- Force 11: Joint Declaration of Data Citation Principles, <https://www.force11.org/datacitation/>
- Gibney E. & Van Noorden R. (2013): Scientists losing data at a rapid rate. Nature, 19.12.2013,
- Goodman A., Pepe A., Blocker A. W. et al. (2014): Ten Simple Rules for the Care and Feeding of Scientific Data. In: PLoS Computational Biology 10(4): e1003542,

References and Further Resources

- Guidelines for the deposit and preservation of research data in Canada
- <https://www.rdc-drc.ca/wp-content/uploads/Guidelines-for-Deposit-of-Research-Data-in-Canada-2015.pdf>
- NIH Data Sharing Policy and Implementation Guidance https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
- OpenAIRE: Open access to research data: the Open Research Data Pilot, <https://www.openaire.eu/h2020-oa-data-pilot>
- Outputs of the NISO Alternative Assessment Metrics Project <https://t.co/E5fY8zYb8G>
- PARSE. Insight: Insight into digital preservation of research output in Europe. Project deliverable D3.4: Survey Report, 9 December 2009, http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

References and Further Resources

- Science magazine: Science Staff introduction to the Special Issue “Dealing with Data”, Science, Vol. 331 no. 6018, 11 February 2011, pp. 692-693, <http://www.sciencemag.org/content/331/6018/692.short>
- The National Science Foundation (NSF) required a data management plan (DMP) <https://www.nsf.gov/bio/pubs/BIODMP061511.pdf>
- Wellcome Trust. (1996). Summary of principles. Retrieved from <http://www.gene.ucl.ac.uk/hugo/bermuda.htm>

Thank you