Northwestern Medicine®

Sharing and Reusing Research Data

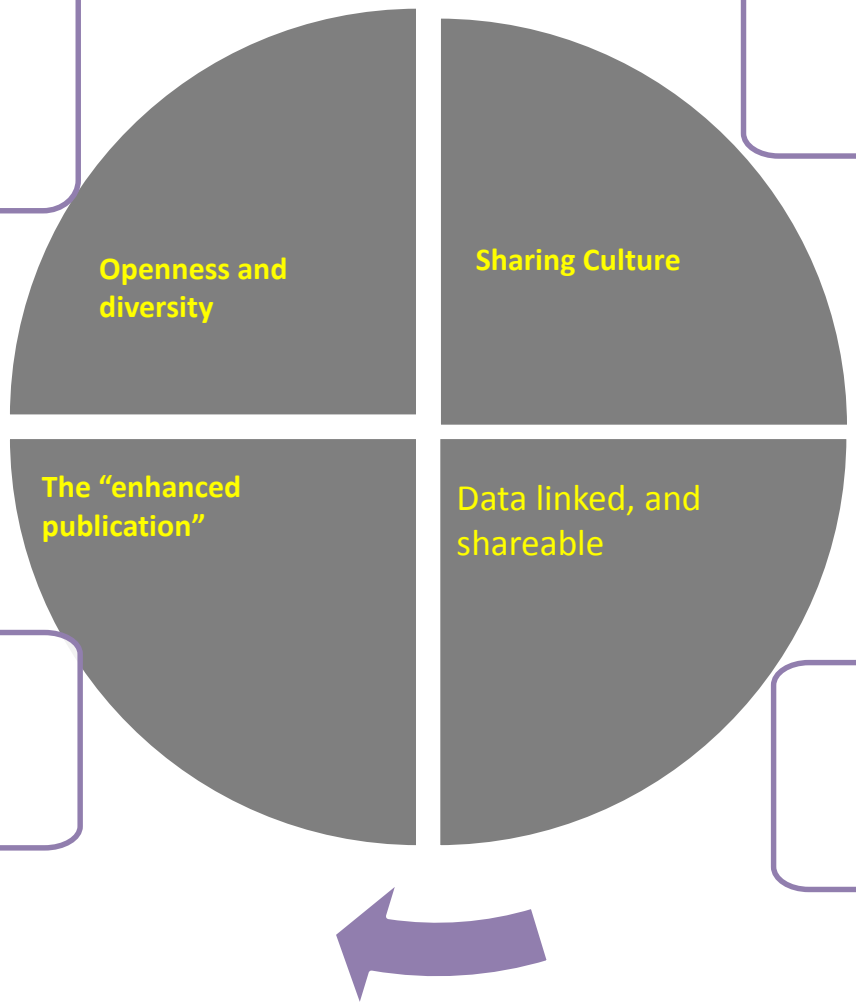
Ehsan Mohammadi, PhD

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



Translation and engagement

- Lectures
- Presentations

Conceptualisation

- Literature reviews
- Bibliographies

Findings

- Books
- Journal articles

Data collection

- Datasets
- Recorded interviews

Northwestern Medicine®

https://goo.gl/AKEHYQ

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

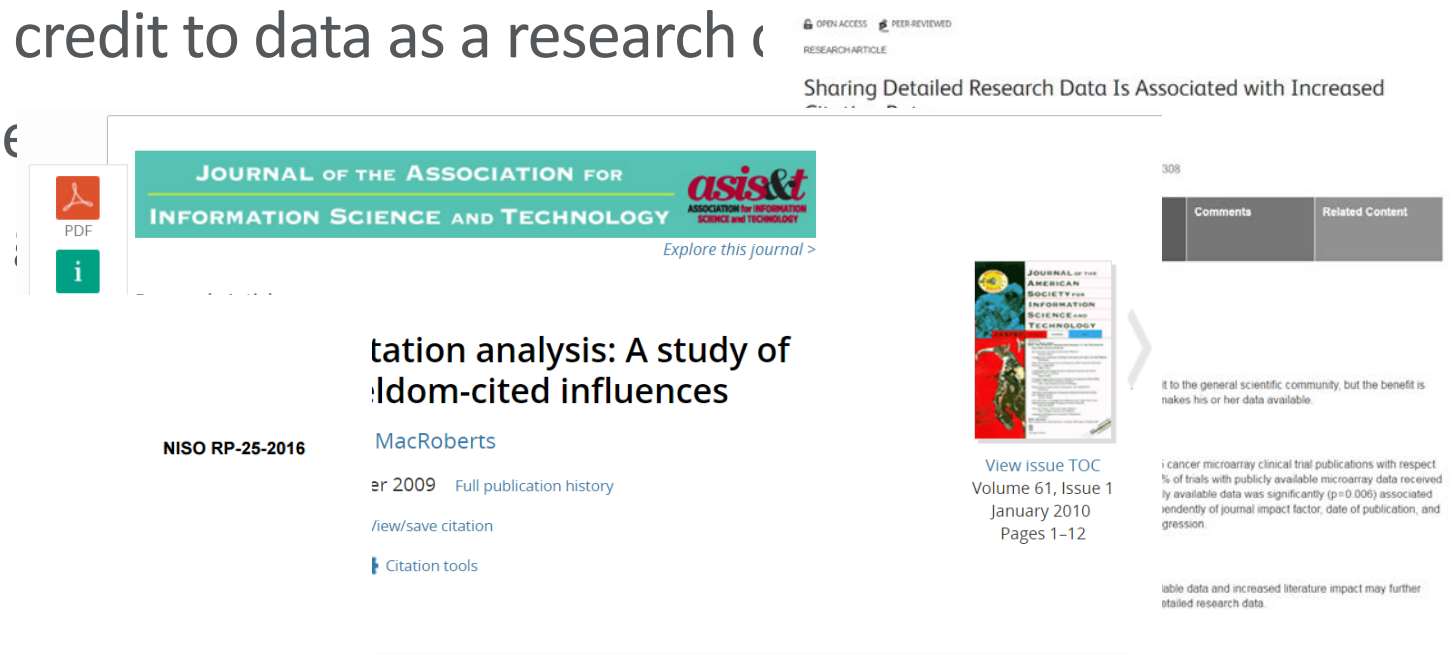
Data can be used and re-used for future research, if:

- ✓ 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
- The succe
- Provides



Outputs of the NISO Alternative Assessment Metrics Project

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

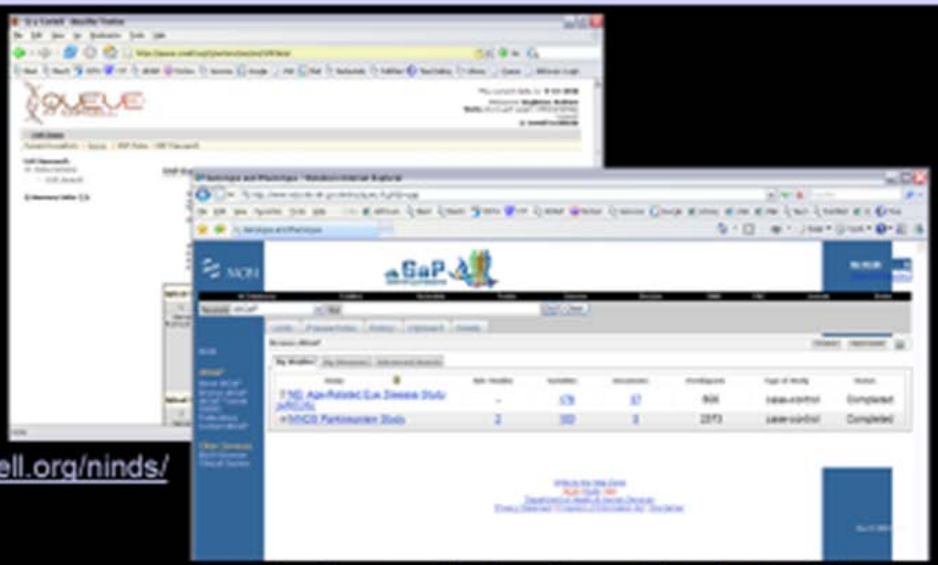
Why share data?

Andrew Singleton
Molecular Genetics Section
Laboratory of Neurogenetics
National Institute on Aging
singleta@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'

https://www.youtube.com/watch?v=Bh6FQjY0NEY

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?

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

The Digital Curation Centre

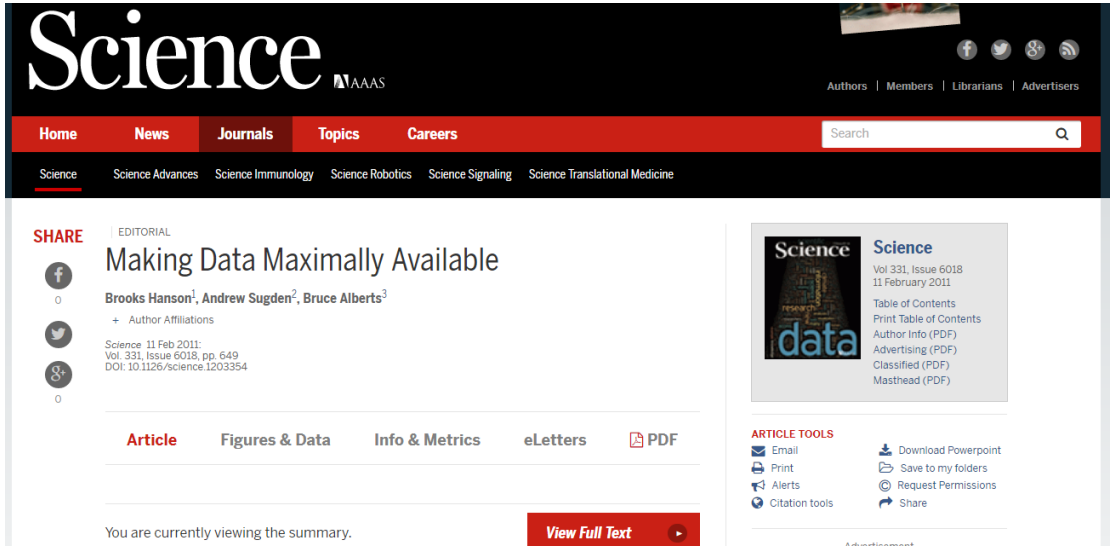
wellcome trust

Australian Government
Australian Research Council

Northwestern Medicine

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



Northwestern Medicine®

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

Northwestern Medicine®

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.


TIME FOR A BREAK

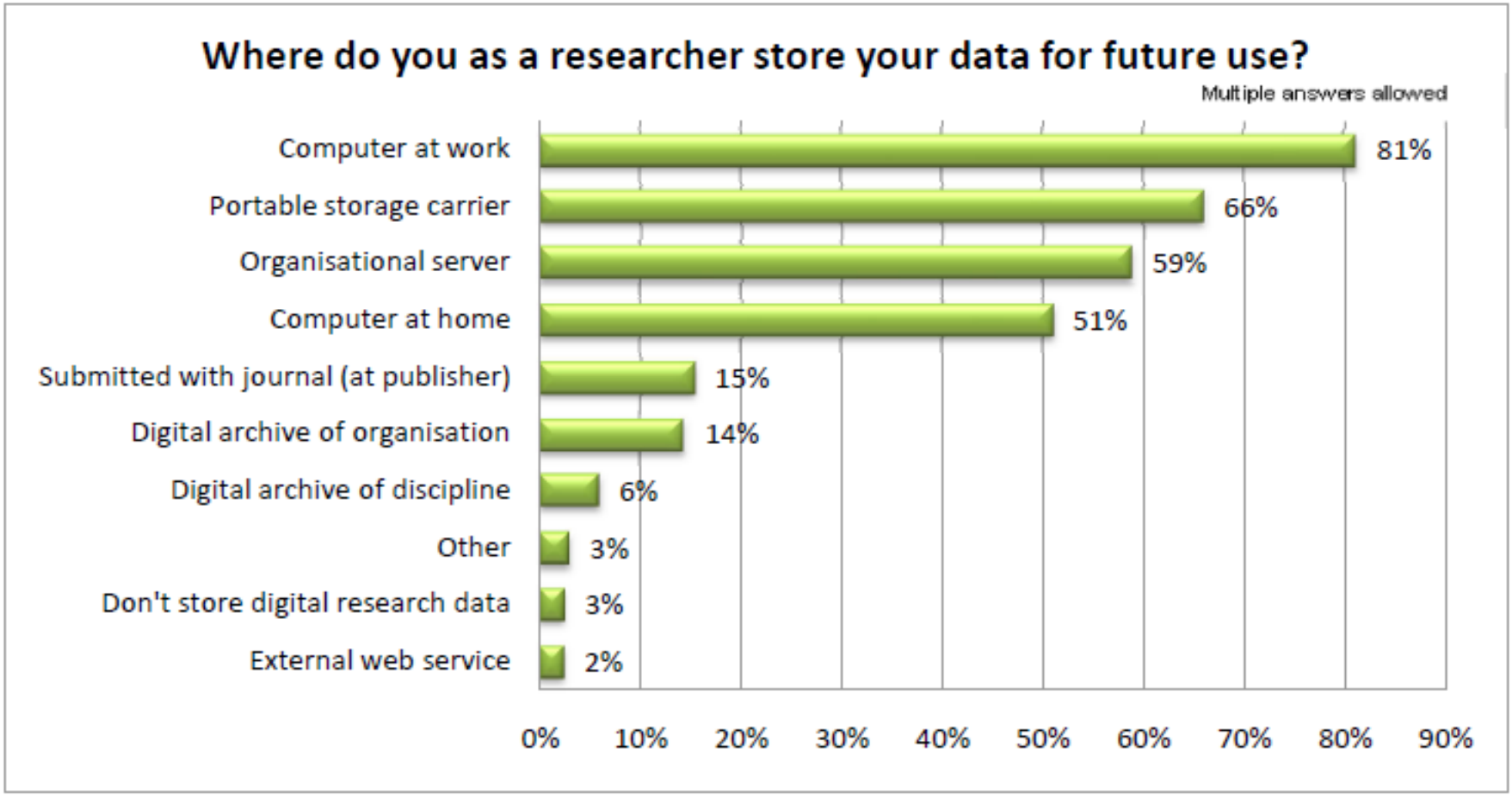
Data sharing barriers

- 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?

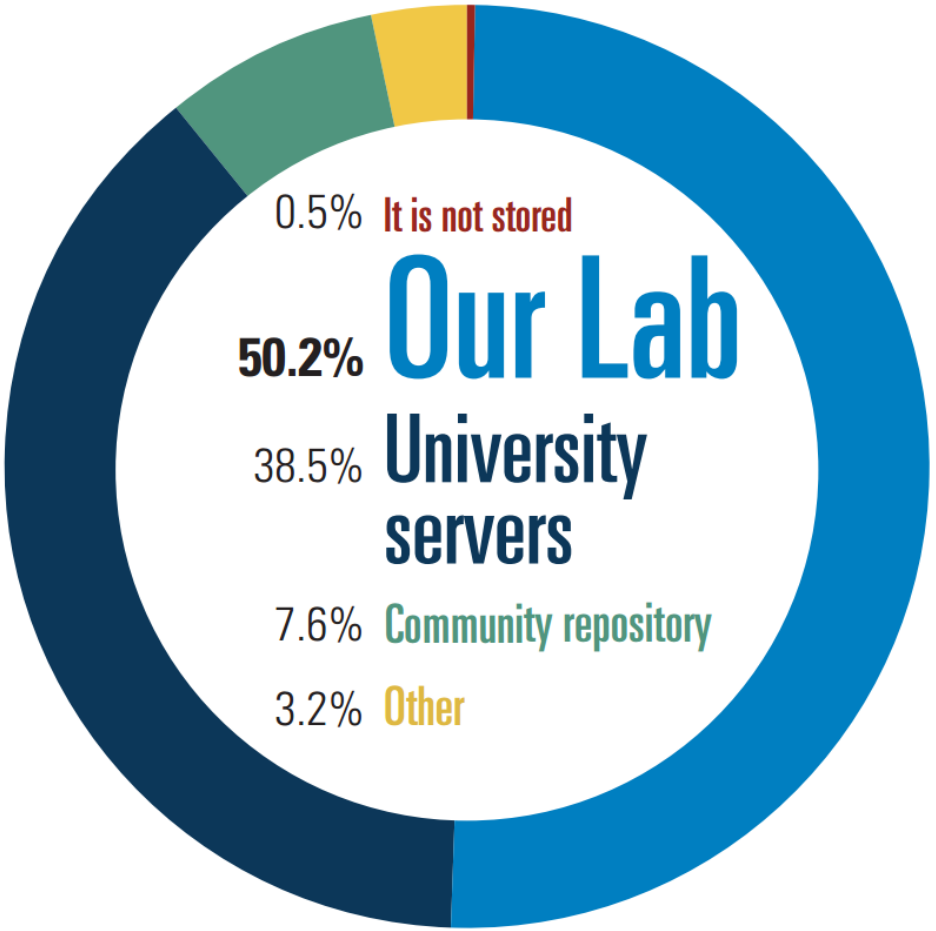


https://goo.gl/057Nv7

Northwestern Medicine®

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

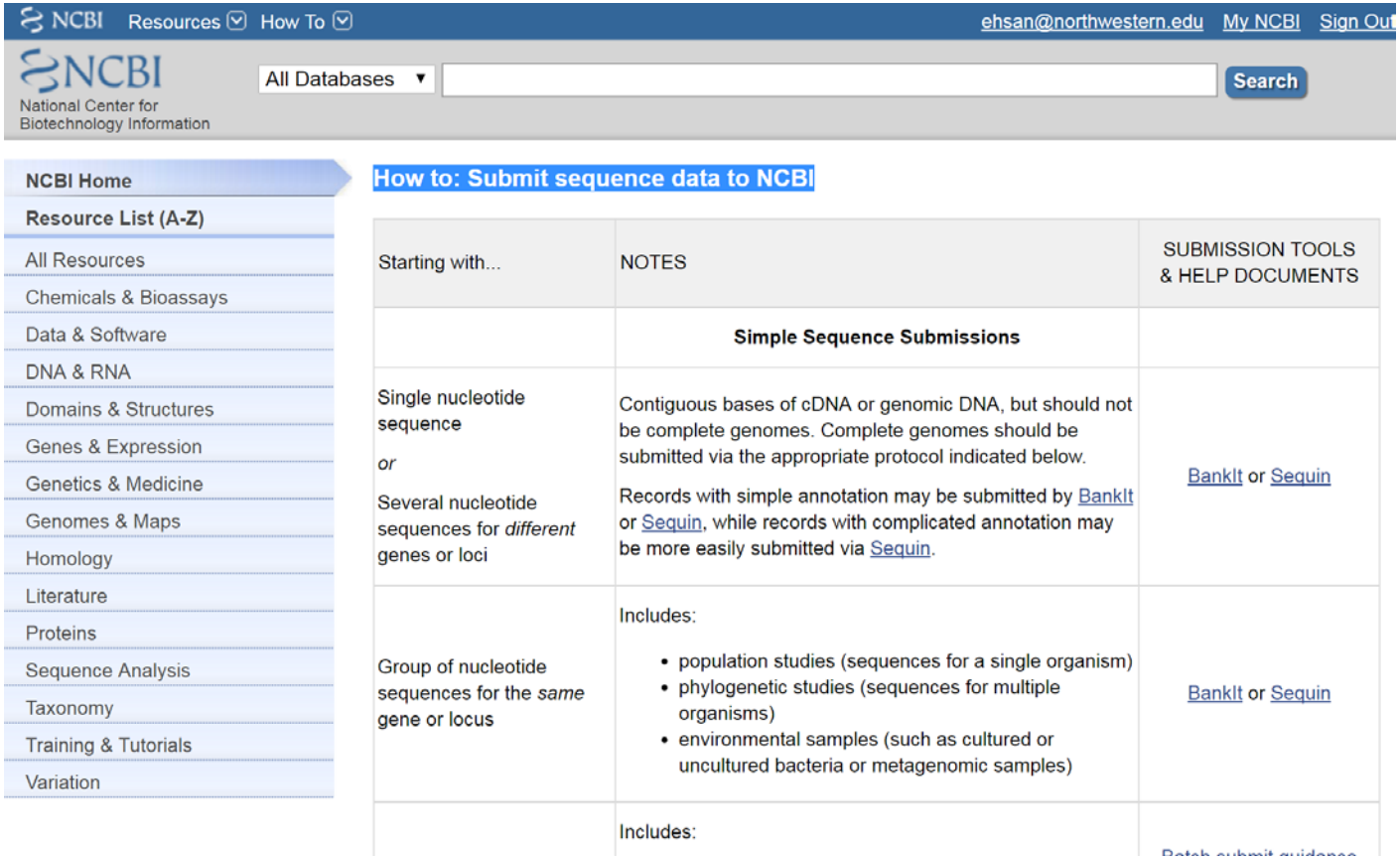
- ✓ Documentation should thoroughly explain the data collection. Examples of documentation files include codebooks, data collection instruments, summary statistics, project summaries, and bibliographies of publications pertaining to the data.

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

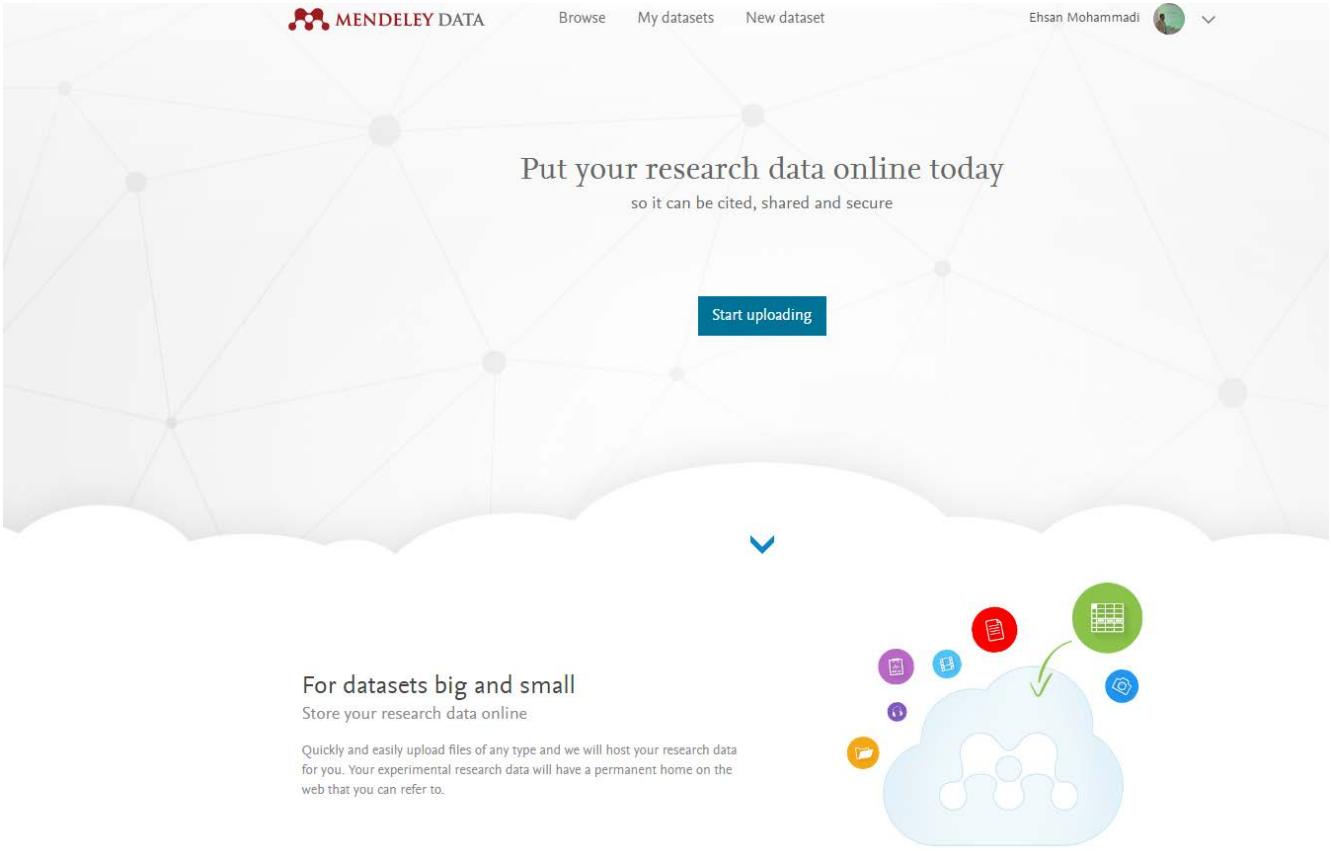
How to: Submit sequence data to NCBI

Starting with...	NOTES	SUBMISSION TOOLS & HELP DOCUMENTS
	Simple Sequence Submissions	
Single nucleotide sequence *or* Several nucleotide sequences for *different* 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 *same* gene or locus	Includes: • 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:	

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

Northwestern Medicine®

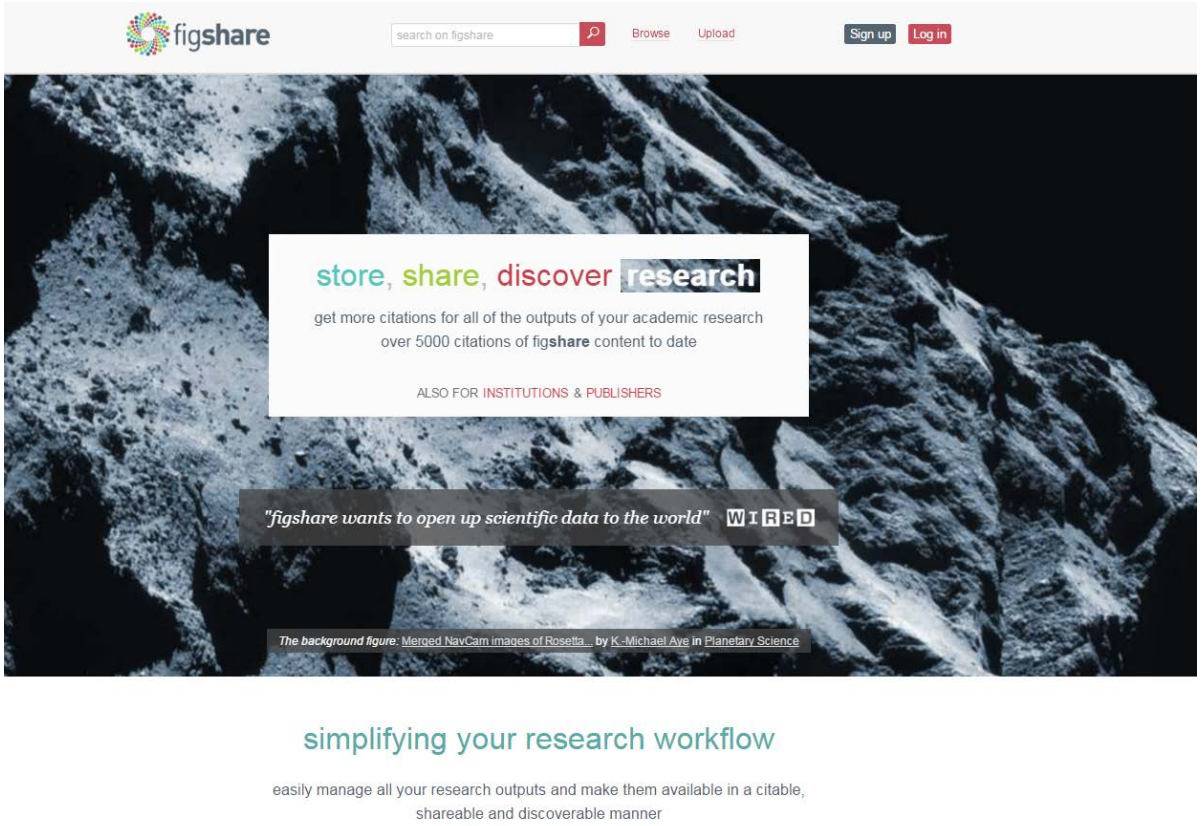
Where should I deposit my data?



https://data.mendeley.com

Northwestern Medicine®

Where should I deposit my data?

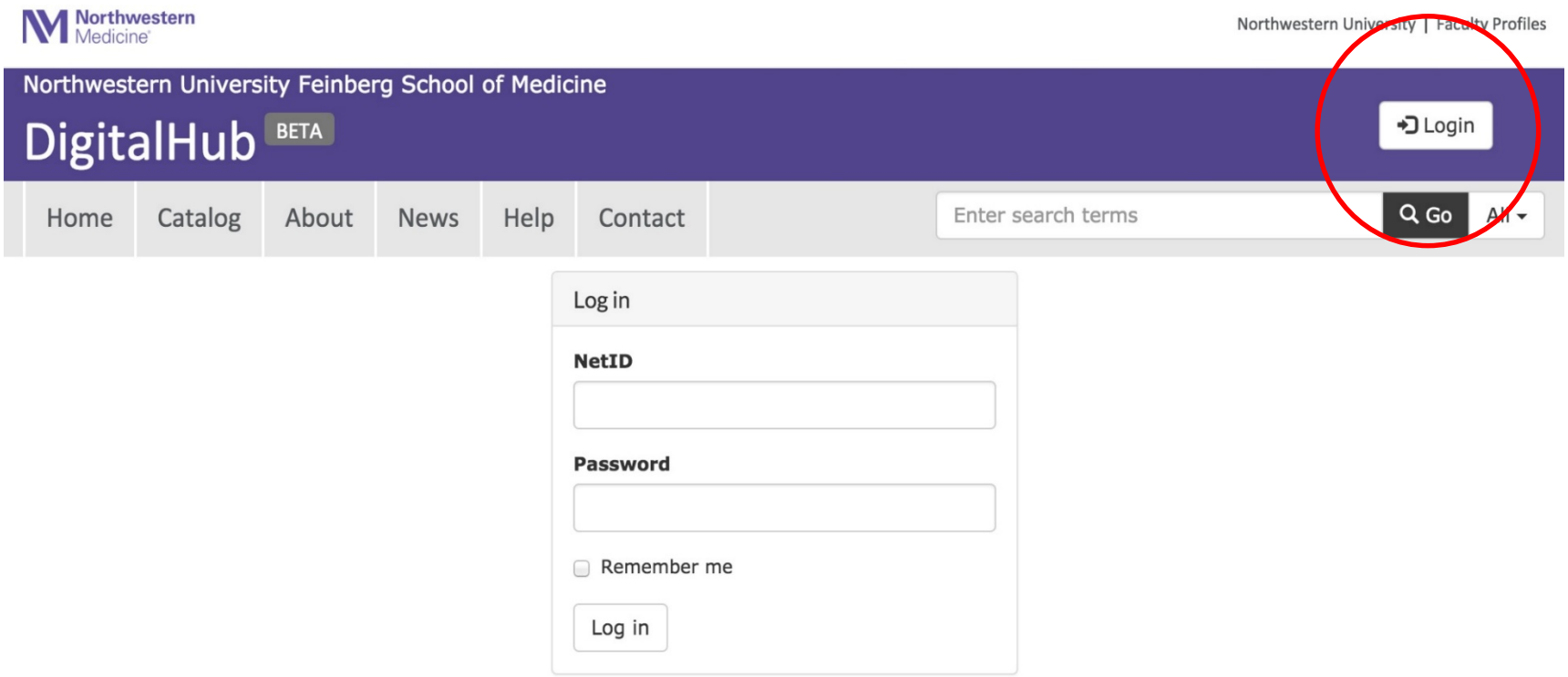


https://figshare.com

Northwestern Medicine®

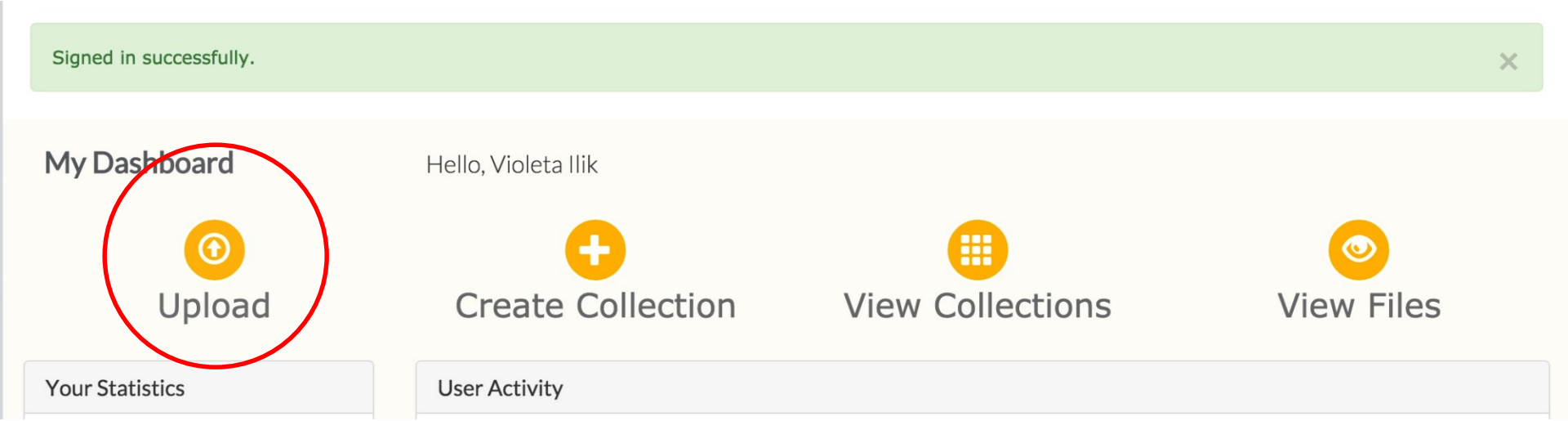
Where should I deposit my data? (DigitalHub)

Step1: Login digitalhub.northwestern.edu



Where should I deposit my data? (DigitalHub)

Step2: Upload

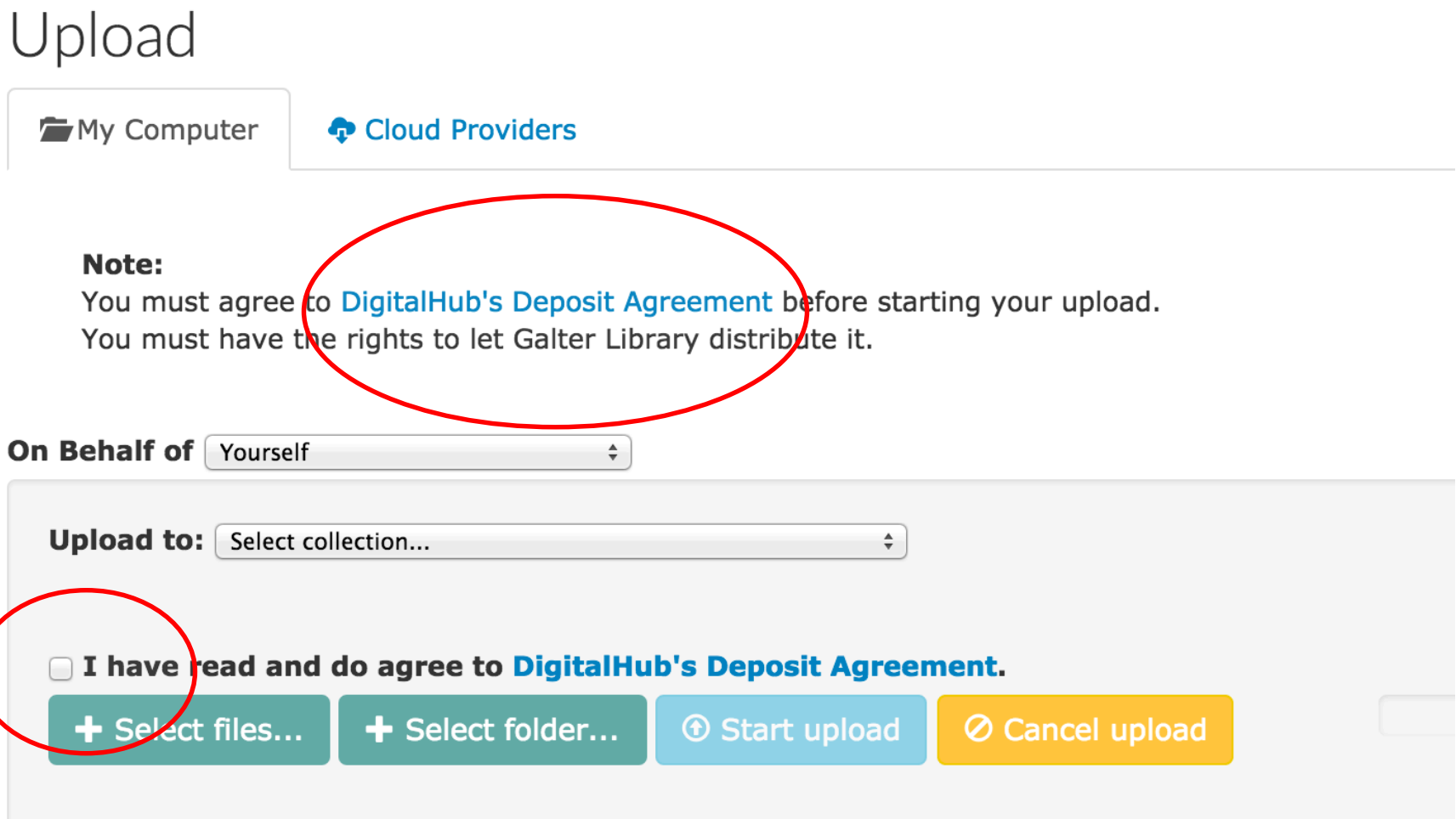


digitalhub.northwestern.edu/dashboard

Northwestern Medicine®

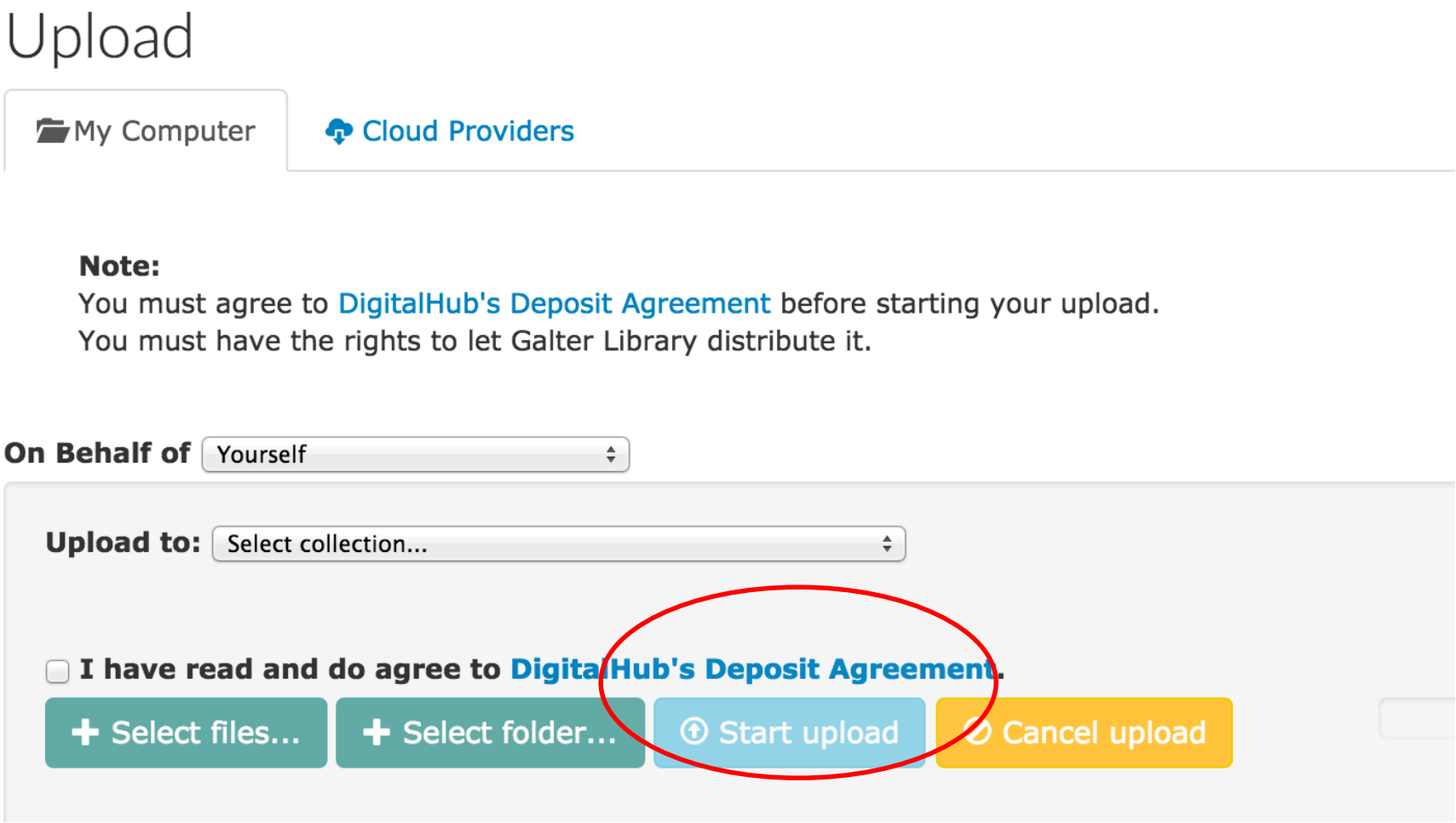
Where should I deposit my data? (DigitalHub)

Step3: deposit agreement



Where should I deposit my data? (DigitalHub)

Step4: Start Upload



Where should I deposit my data? (DigitalHub)

Step5: Apply Metadata

5 required fields:

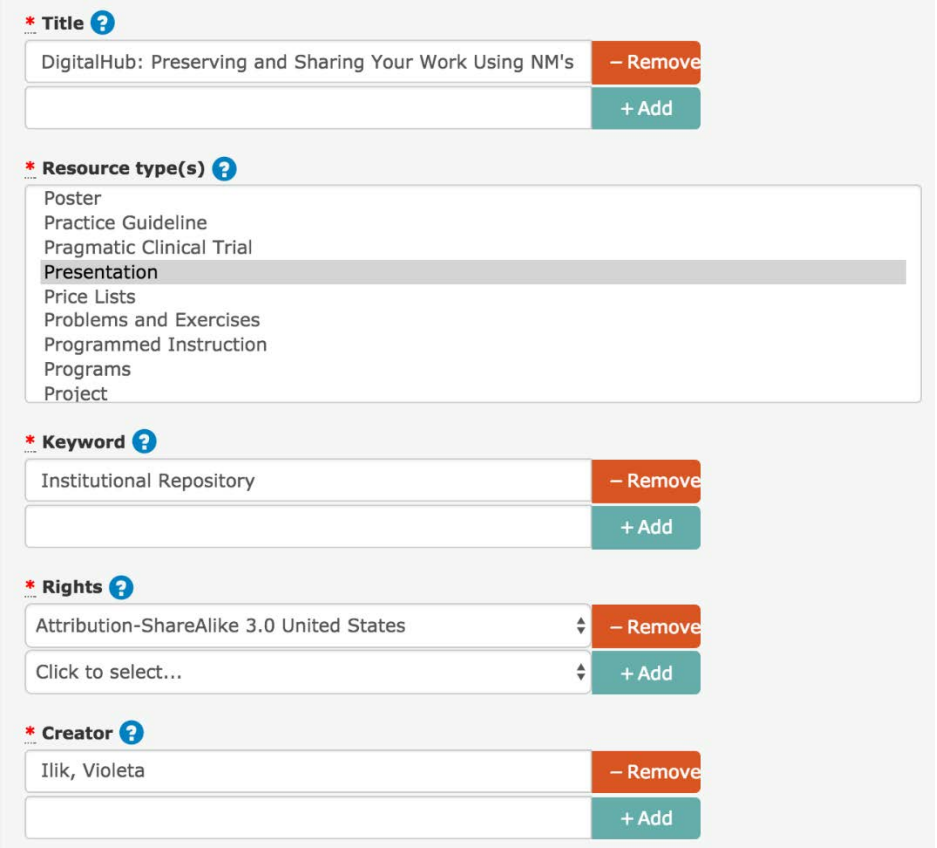
Title

Resource type

Keyword

Creator

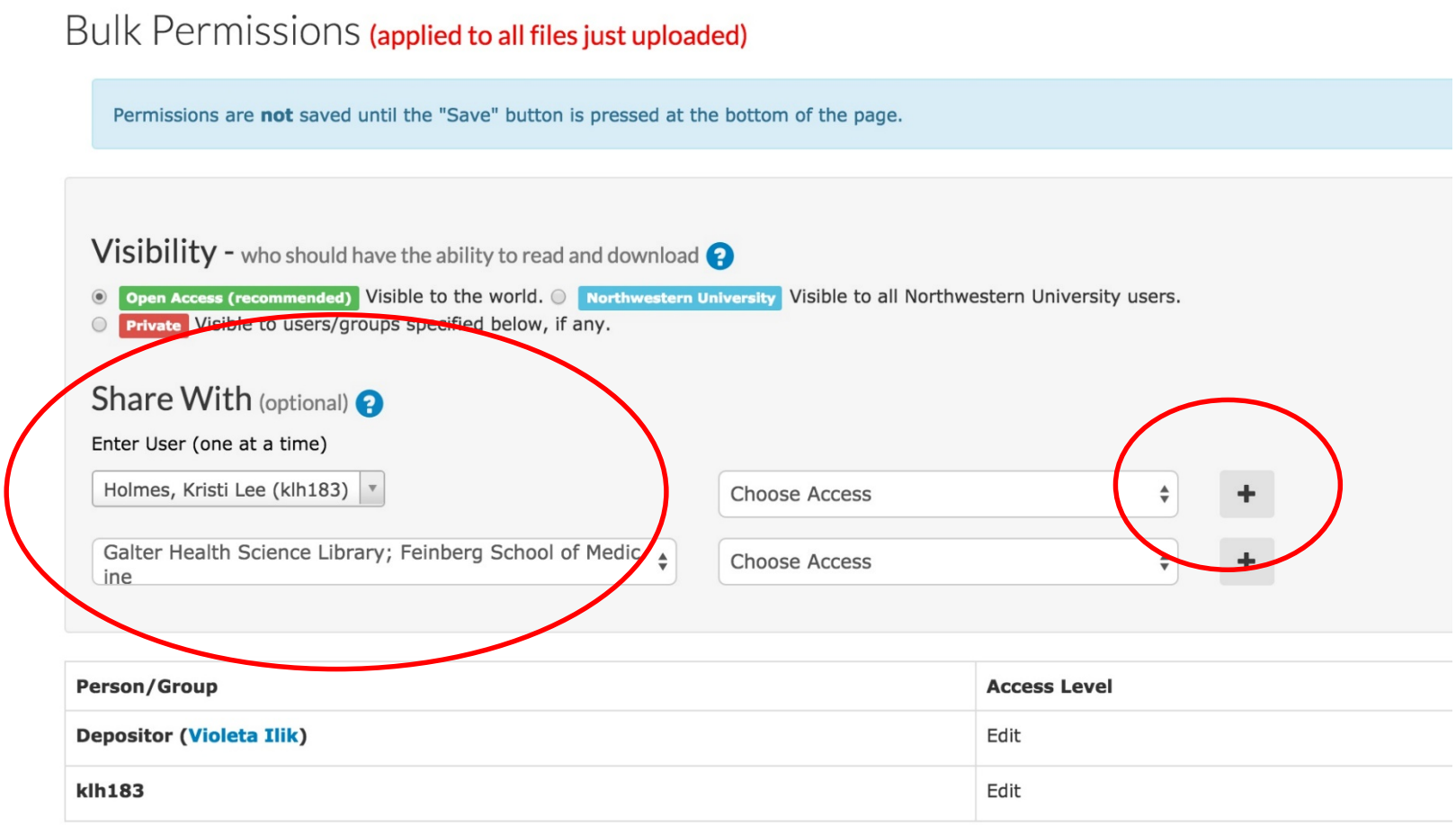
Rights



Northwestern Medicine®

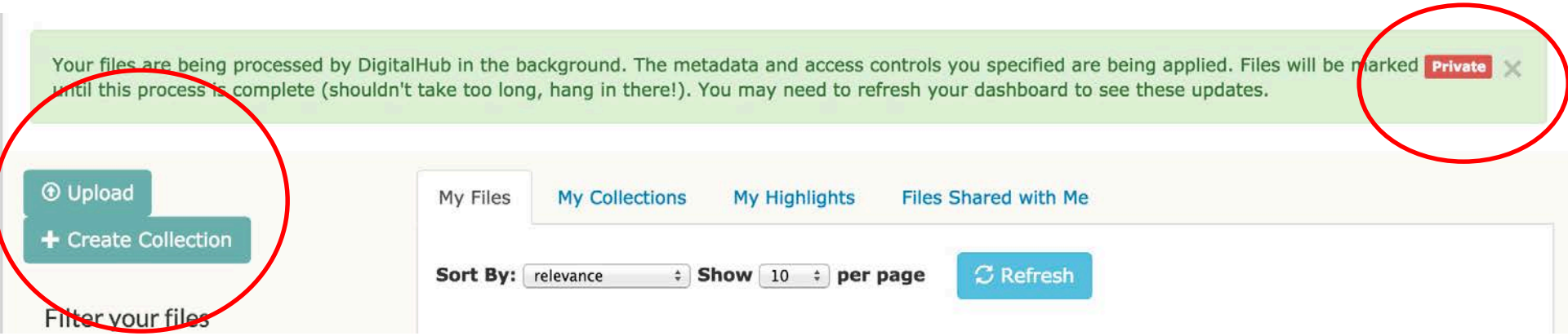
Where should I deposit my data? (DigitalHub)

Step6: visibility & permissions



Where should I deposit my data? (DigitalHub)

Step7: Save!



Congratulations, you have submitted your work!

Northwestern Medicine®

How to find data for future studies?

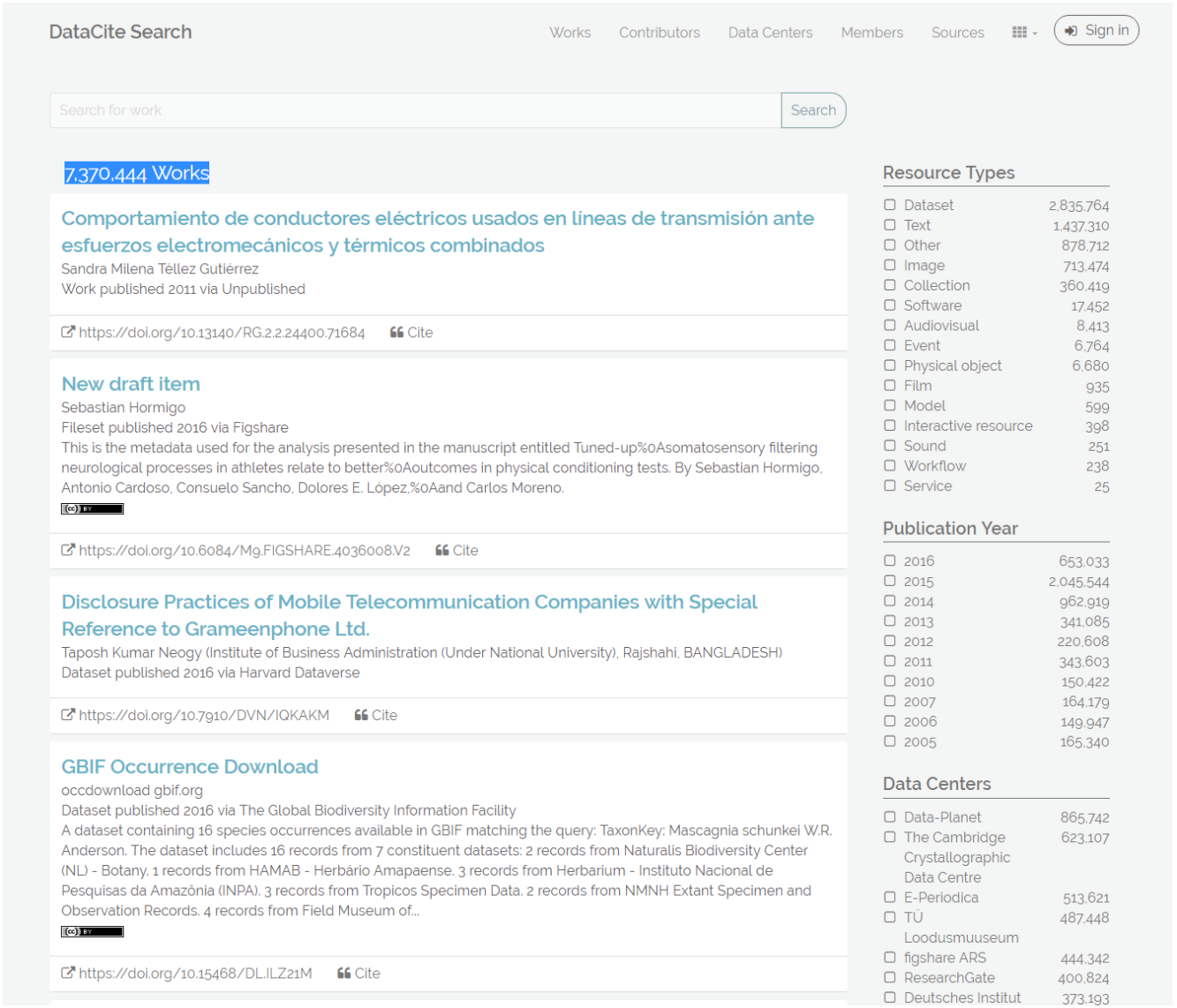
Subject repositories

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/



Northwestern Medicine®

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

https://www.nlm.nih.gov/NIHbmic/data_sharing_plan.html

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=Bh6FQjY0NEY

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 Guidancehttps://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