

Research Data Management Support at Cornell

<https://data.research.cornell.edu>

rdmsg-help@cornell.edu

Wendy Kozlowski (wak57), RDMSG Coordinator

What data management IS:

(v.) Action that contributes to effective storage, preservation and use of data and documentation throughout the research lifecycle

What data management is NOT:

- Data or computation science
- Database administration
- A research method
 - What data to collect
 - How to collect data
 - How to design an experiment

Why we want researchers to have good data management practices

- Reproducibility
(Journals)
- Recognition
(Researchers)
- Reuse
(Funders)



"You can't keep coming in here and demanding data every two years!"

What is data management plan (DMP)?

In the context of research funders, a DMP is a document outlining a strategy for storage, preservation and sharing (reuse) of data related to a specific project.

Who requires a DMP-like document with a proposal?

Key Cornell Funders

- [NSF](#) DMP
- [NIH](#) Data Sharing Plan
- USDA/[NIFA](#) DMP
- [NOAA](#) DMP
- [NASA](#) DMP
- [USGS](#) DMP
- [DOE](#) DMP

Other Funders/More Information

- Gordon and Betty Moore Foundation
- Gulf of Mexico Research Institute
- IMLS
- NEH/Office of Digital Humanities
- Smithsonian Institute Digital Asset Management Plan
- <https://data.research.cornell.edu/content/funder-data-requirements>
- <http://datasharing.sparcopen.org/>



Office of Budget Finance & Award Management (BFA)

Office of Budget, Finance, & Award Management >

Budget Division >

Division of Acquisition and Cooperative Support >

Division of Financial Management >

Division of Grants & Agreements >

Division of Institution & Award Support

CAAR Branch

Policy Office

Systems Office

Staff

Large Facilities Office >

Advisory Committees

Home > Budget Finance & Award Manage... > Institution and Award Support

Email Print Share

Dissemination and Sharing of Research Results

NSF DATA SHARING POLICY

Since at least 03 Aug 1995

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 [Proposal & Award Policies & Procedures Guide \(PAPPG\) Chapter XI.D.4.](#)

NSF DATA MANAGEMENT PLAN REQUIREMENTS

Announced 01 Oct 2010

Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. See [PAPPG Chapter II.C.2.j](#) for full policy implementation.

REQUIREMENTS BY DIRECTORATE, OFFICE, DIVISION, PROGRAM, OR OTHER NSF UNIT

Links to data management requirements and plans relevant to specific Directorates, Offices, Divisions, Programs, or other NSF units, are provided below. If guidance specific to the program is not provided, then the requirements established in [PAPPG Chapter II.C.2.j](#) apply.



Policy & Compliance

[NIH Grants Policy Statement](#)

[Notices of Policy Changes](#)

[Compliance & Oversight](#)

[Select Policy Topics](#)

+

NIH Data Sharing Policy

Data sharing is essential for expedited translation of research results into knowledge, products and procedures to improve human health.

2003

The [Final NIH Statement on Sharing Research Data](#) was published in the NIH Guide on February 26, 2003. This is an extension of NIH policy on sharing research resources, and reaffirms NIH support for the concept of data sharing. The new policy becomes effective with the October 1, 2003 receipt date for applications or proposals to NIH.

- [Data Sharing Regulations/Policy/Guidance Chart for NIH Awards \(08/30/2006\)](#) - (MS Word - 58 KB) - This chart is designed as a quick guide only for the purpose of identifying various data sharing regulation/policy/guidance documents applicable to NIH funding.
- [NIH Guide Notice \(02/26/2003\)](#) - Final NIH Statement on Sharing Research Data.
- [NIH Guide Notice \(03/01/2002\)](#) - NIH Announces a Draft Statement on Sharing Research Data.
- [NIH Data Sharing Policy and Implementation Guidance \(03/05/2003\)](#) - Guidance providing the NIH policy statement on data sharing and additional information on the implementation of this policy.
- [Frequently Asked Questions - Data Sharing \(02/16/2004\)](#) - Listing of Frequently Asked Questions that will be updated as new questions are received. Please check back periodically for new questions and answers.
- [Data Sharing Workbook \(PDF - 75 KB\) or \(MS Word - 74 KB\) - \(02/16/2004\)](#) - Workbook to show how investigators working in a variety of scientific areas have shared their data.
- [NIH Data Sharing Brochure \(PDF - 244 KB\) - \(05/20/2003\)](#) - Printable brochure that summarizes main elements of the NIH Data Sharing Policy.
- [Testimonials \(MS Word - 22 KB\) - \(03/05/2003\)](#) - First-hand accounts from researchers who have shared data.
- [Other Data Sharing Documents and Resources \(02/19/2004\)](#) - Additional resources relating to data sharing.
- [For NIH Staff Use - \(02/16/2004\)](#)

Proposals with \geq \$500,000/year
Direct Costs requires
Data Sharing Document



NIH Genomic Data Sharing

To set forth expectations that ensure the broad and responsible sharing of genomic research data, NIH issued the Genomic Data Sharing (GDS) Policy in August 2014, both the *NIH Guide Grants and Contracts* (available at <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html>), and the *Federal Register* (available at <https://federalregister.gov/a/2014-20385>).

The NIH GDS Policy became effective for competing grant applications submitted for the January 25, 2015, receipt date; contract proposals submitted to NIH on or after January 25, 2015; and for intramural projects generating genomic data on or after August 31, 2015. The NIH GDS Policy applies to all NIH-funded research (e.g., grants, contracts, and intramural research) that generates large-scale human or non-human genomic data, regardless of the funding level, as well as the use of these data for subsequent research. Large-scale data include genome-wide association studies (GWAS), single nucleotide polymorphisms (SNP) arrays, and genome sequence, transcriptomic, epigenomic, and gene expression data. Examples of genomic research projects that are subject to the Policy and the timeline for submission and sharing of data from such projects may be found in the Supplemental Information to the NIH GDS Policy available at the NIH GDS Policies link below.

Questions about the Policy can be e-mailed to GDS@mail.nih.gov.

- [NIH GDS Policy](#)
- [NIH GDS Policy Oversight](#)
- [Guidance for Submitting and Requesting Access to Controlled-Access Data Maintained in an NIH-Designated Data Repository \(e.g., dbGaP\)](#)
 - [Study Registration and Data Submission](#)
 - [Institutional Certifications](#)
 - [Request to Access Data](#)
- [Data Repositories and NIH Trusted Partners](#)
- [Related Resources](#)
- [Facts and Figures](#)

Scientific Data Sharing

- > [Genomics and Health](#)
- > [Scientific Data Management](#)

Frequently Asked Questions

- > [NIH Genomic Data Sharing Policy](#)

Key DMP Takeaways

- Funders are looking to know what data and metadata is going to be generated, how it will be kept secure and if, when and how it will be shared
- For cross-institutional proposals, most funders say data management is ultimately the responsibility of the lead-PI, and a proposal will require just one DMP.
- Be sure to read RFP/RFI calls carefully – some directorates/divisions/offices/programs have specific requirements that need to be addressed.
- Some funders consider them part of the peer-review portion, others consider them only once a project is funded.
- Don't wait until the last minute to write them – they DO matter!

How to get help with Data Management Plans

- Agency guidance (see links provided throughout)
- RDMSG guidance <https://data.research.cornell.edu>
 - [Data management planning](#)
 - [Data storage finder tool](#)
 - [Data management best practices](#)
- DMPTool: Web-based tool to build and edit a customized plan according to select funder requirements. <https://dmptool.org>
- Ask for help! The RDMSG supports Cornell researchers and provides free consultations. rdmsg-help@cornell.edu

Publisher Data Sharing Requirements

<http://journals.plos.org/plosone/s/data-availability>



Data Availability

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.

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

Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection.

<http://www.sciencemag.org/authors/science-journals-editorial-policies>



Data and materials availability after publication

After publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of a *Science Journal*. After publication, all reasonable requests for data, code, or materials must be fulfilled. Any restrictions on the availability of data, code, or materials, including fees and restrictions on original data obtained from other sources must be disclosed to the editors...

Unreasonable restrictions on data, code, or material availability may preclude publication.

<http://www.nature.com/authors/policies/availability.html>

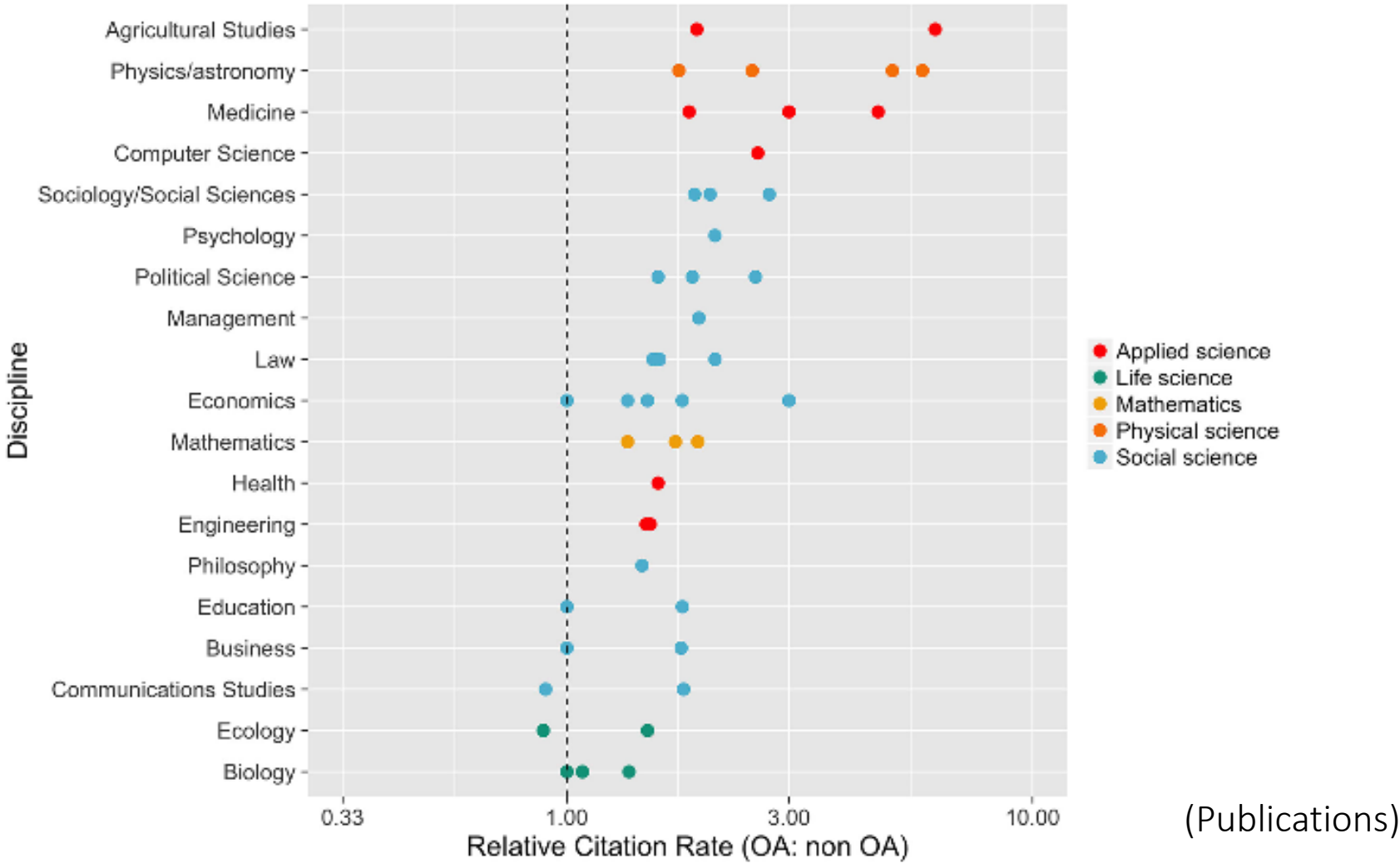


Availability of data, materials and methods

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 at the time of submission. Any restrictions must also be disclosed in the submitted manuscript.

Supporting data must be made available to editors and peer-reviewers at the time of submission for the purposes of evaluating the manuscript. ... The preferred way to share large data sets is via public repositories.

Researcher Recognition



Researcher Recognition

Increased citation rate for studies with openly share datasets

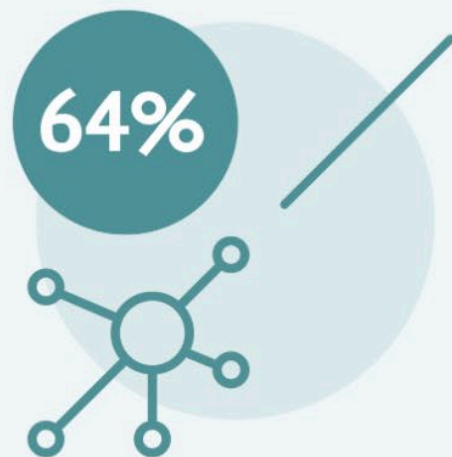
Discipline	Rate of increased citation	Source
Gene expression microarray data	9% overall, 30% for older studies	Piowar and Vision, 2013
Astronomy	20%	Henneken and Accomazzi, 2011
Astrophysics	28-50%	Dorsch et al, 2013
Paleoceanography	35%	Sears, 2011

(Data)

Data Sharing - Perceptions



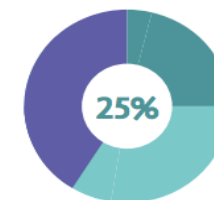
73% of academics surveyed said that having access to published research data would benefit their own research



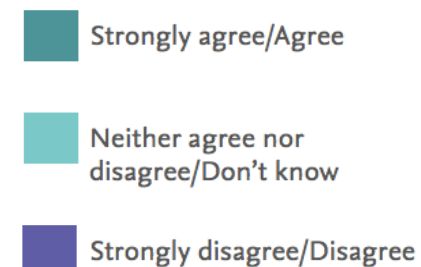
64% are willing to allow others to access their research data



69% of survey respondents said sharing research data is important for doing research in their field



I have received sufficient training in research data sharing





Sharing data at Cornell

<https://ecommons.cornell.edu>

Available to Cornell affiliates and provides:

- Open access
- Persistent identifiers (handle, DOI)
- Flexible, optional licenses
- Curatorial review
- Links to and citations for related material
- Page and file analytics
- Long term sharing and preservation

Cornell University

eCommons

Cornell's digital repository

Search

Browse - About - Help Login

eCommons is a service of Cornell University Library that provides long-term access to a broad range of Cornell-related digital content of enduring value. [Learn more](#)

Make a deposit
Submit your document, research paper, image, data set and more into Cornell's digital repository.
[Submit your work](#) [How to submit](#)

49,431 items in eCommons

✓ 7,661 Theses and Dissertations ✓ 1,936 Images ✓ 817 Videos ✓ 3,372 Technical Reports ✓ 134 Data Sets [View more](#)

Last month's top downloads

[Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications](#)
(5,056 downloads)
[Mission, Vision, Values & Goals](#)

Recently added

[CVM eNews, December 2018](#)
Office of Marketing and Communications

[NEVBD Quarterly Digest - December 2018](#)
Northeast Regional Center for Excellence in Vector-Borne Diseases

FAIR DATA

DATA SHOULD BE

Findable

All Data Objects should be uniquely and persistently identifiable from other data types.

Accessible

Data can be accessible machines and human users, through well-defined protocols and appropriate authorization.

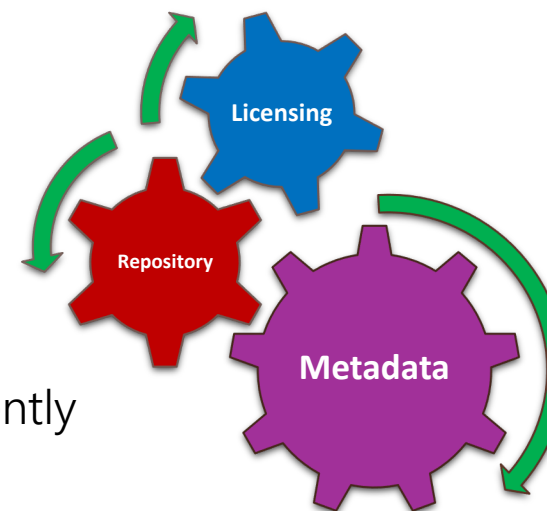
Interoperable

Metadata are machine-actionable in automated workflows, syntactically parseable, and utilize shared vocabularies within and across domains.

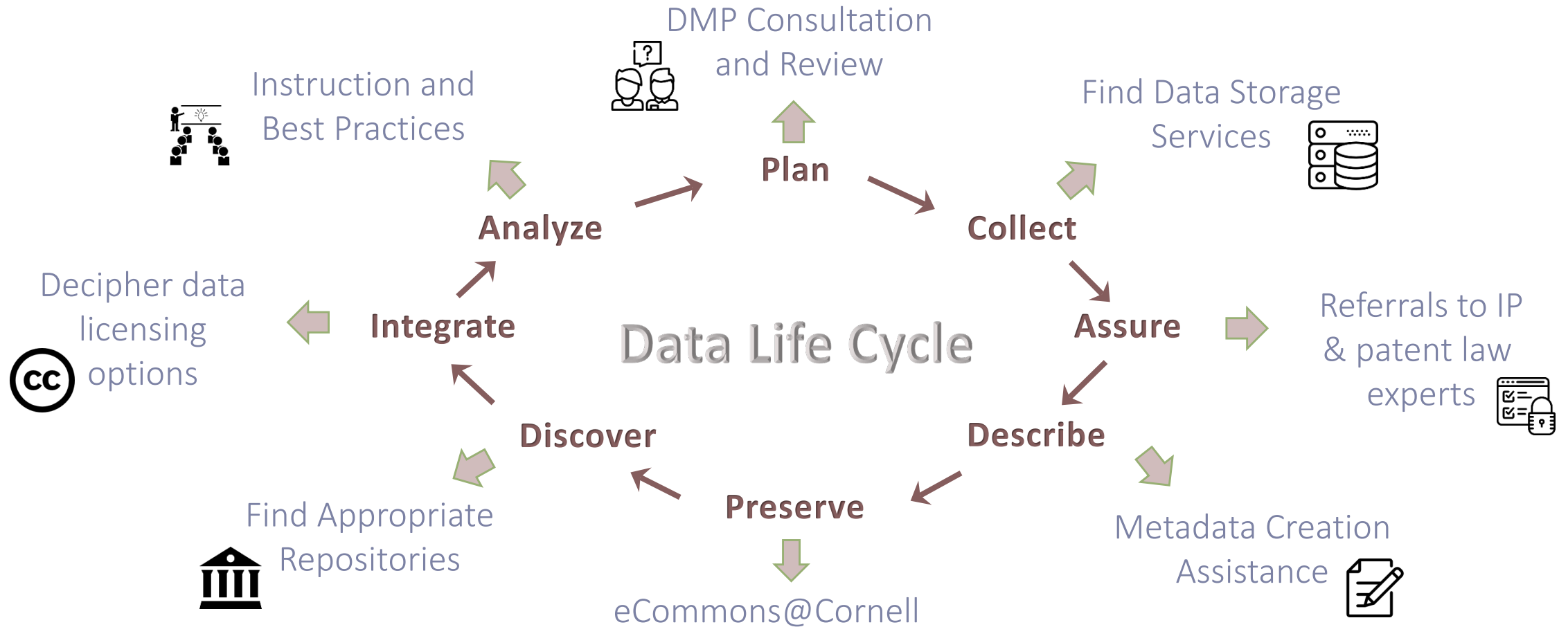
Reusable

Data are sufficiently well-described to be automatically linked and/or integrated, with sufficient provenance and citation.

BY HUMANS AND MACHINES



Data Management Services at Cornell



✉ rdmsg-help@cornell.edu
🌐 data.research.cornell.edu



Data Storage Finder

1. What is the classification of your data? ⓘ

- Public
- Sensitive / Moderate Risk
- Confidential or Restricted / High Risk
- HIPAA-Regulated

2. Do you need backups, snapshots or replication of your data? ⓘ

- I need one or more backup/snapshot copies of the data, and need to be able to restore data from previous points in time (high durability).
- I need to have replicate copies of the data to minimize downtime (high availability).

3. How much data do you have and how fast will it grow? ⓘ

- Unlikely to exceed 1TB in 2 years
- Greater than 1TB or likely to exceed in 2 years

First:

Describe your data

4. Do you have special performance needs? ⓘ

- I am likely to have more than 1,000 files in a single directory within two years.
- My data interactions demand high transaction or transfer rates.

5. How are you expecting to access the data? ⓘ

- I need easy access to this data from anywhere, even when I don't have my own computer or mobile device with me.
- I frequently need access from a mobile device such as a phone or tablet.

6. With whom do you need to share your data regularly? ⓘ

- Only those with a Cornell NetID or GuestID
- Only those with a Weill CWID
- Users in and out of the Cornell community



Second:

Data Storage Finder

→ Select data storage services you would like to compare.









<p>Amazon Web Services Elastic Block Store</p> <p>Storage for use with Amazon EC2</p> <input type="radio"/>	<p>Amazon Web Services Elastic File System</p> <p>Storage for use with multiple Amazon EC2 instances</p> <input type="radio"/>	<p>Amazon Web Services Glacier</p> <p>Cloud based archival storage</p> <input type="radio"/>	<p>Amazon Web Services S3</p> <p>Flexible, scalable object storage</p> <input checked="" type="radio"/>	<p>BioHPC Cloud</p> <p>Storage for BioHPC lab computing services</p> <input type="radio"/>
<p>CAC Archival Storage</p> <p>Single copy, non-mountable storage</p> <input type="radio"/>	<p>CAC Red Cloud Storage</p> <p>Storage for Red Cloud compute instances</p> <input type="radio"/>	<p>CISER Research Servers</p> <p>Storage for CISER computing environment</p> <input type="radio"/>	<p>CUGIR</p> <p>Publicly shared geospatial data storage</p> <input type="radio"/>	<p>CUL eCommons</p> <p>Publicly shared data repository</p> <input type="radio"/>
<p>Cornell Box</p> <p>Online file sharing and collaboration</p> <input checked="" type="radio"/>	<p>Cornell Restricted Access Data Center</p> <p>Storage for CRADC (confidential) computing environment</p> <input type="radio"/>	<p>EZ-Backup Static Storage</p> <p>Archival storage and backup storage</p> <input type="radio"/>	<p>Google Drive</p> <p>Cornell G Suite file storage and sharing</p> <input type="radio"/>	<p>Kaltura Video Platform Service</p> <p>Flexible, scalable video and multi-media storage (customizable)</p> <input type="radio"/>
<p>Kaltura Video on Demand</p> <p>Video and multi-media storage</p> <input type="radio"/>	<p>LabArchives</p> <p>Online electronic lab notebook</p> <input type="radio"/>	<p>Open Science Framework</p> <p>Online project management repository</p> <input type="radio"/>	<p>Shared File Services</p> <p>File sharing between users and computers</p> <input checked="" type="radio"/>	<p>Shared File Services - Confidential</p> <p>File sharing between users and computers for confidential data (non-WCM)</p> <input type="radio"/>
<p>WCM Block Storage</p> <p>High performance storage attached to centrally hosted servers (WCM only)</p> <input type="radio"/>	<p>WCM File Sharing</p> <p>Secure network storage (NFS/CIFS) for research computing (WCM only)</p> <input type="radio"/>	<p>WCM Red Cloud Secure Storage</p> <p>Secure storage for Red Cloud compute instances (WCM only)</p> <input type="radio"/>	<p>WCM Secure Remote Archive</p> <p>Secure, single copy, non-mountable storage (WCM only)</p> <input type="radio"/>	



Data Storage Finder

Third:

Compare services that match your selected criteria

Cost ⓘ	 -  up to \$1000/TB/Year Cost dependent on storage class, replication and other services used.	 Free for Cornell users	  \$500-\$1000/TB/Year Cost dependent on replication configuration.
Capacity ⓘ	5 TB file size limit. No overall limit (costs incurred). No practical limit to number of files.	15 GB file size limit. No overall limit. No limit to number of files.	No limit (costs incurred).
Data Allowed ⓘ	Allowed: Public data. Sensitive / moderate risk data. Allowed with special configuration: FERPA-protected data. Confidential or restricted / high risk data. Contact Amazon Web Services S3 for more information. Not Allowed: HIPAA-regulated data. Export-Controlled information.	Allowed: Public data. Sensitive / moderate risk data. FERPA-protected data. Not Allowed: Confidential or restricted / high risk data. HIPAA-regulated data. Export-Controlled information.	Allowed: Public data. Sensitive / moderate risk data. FERPA-protected data. Not Allowed: Confidential or restricted / high risk data. HIPAA-regulated data. Export-Controlled information.
Durability (protection against data loss) ⓘ	 High Versioning can be configured to protect against human error (costs incurred).	 High Automatic file versioning keeps 100 most-recent versions.	 High Scheduled snapshots.



Workshops and Instruction

Upcoming Data and GIS workshops at the Libraries

<http://bit.ly/CULibraryWorkshops>

- **Data Documentation: Intro to Science Metadata** | 2019-09-24 | 4:00-5:30 pm | Stone Classroom, Mann Library
- **Data Sharing and Publishing** | 2019-10-01 | 4:00-5:30 pm | Stone Classroom, Mann Library
- **Cleaning Messy Data with OpenRefine** | 2019-10-08 | 4:00-5:30 pm | Stone Classroom, Mann Library
- **Intro to QGIS** | 2019-09-19 | 2:30pm | Stone Classroom, Mann Library
- **Intro to QGIS** | 2019-10-25 | 2:30pm | Stone Classroom, Mann Library
- **Using Excel: Formulas and Functions** | 2019-10-22 | :30-6 pm | Uris B05
- **Using Excel: Pivot Tables** | 2019-10-29 | 4:30-6 pm | Uris B05
- **Using Excel: Macros** | 2019-11-12 | 4:30-6 pm | Uris B05
- **Research Record Keeping: Electronic Lab Notebooks and OSF** | 2019-9-19 | TIME 12-1pm | ILR



- Tuesday October 15, 2019, 9am-5pm
- ILR Conference Center
- More info and agenda: <http://bit.ly/cornelldayofdata2019>
- Registration required for lunch and afternoon workshops
- Keynote by Professor David Mimno
- Researcher panel on data-related challenges and solutions
- Lunch and a data services FAIR
- 4 tracks of concurrent afternoon workshops and informational sessions



✉ rdmsg-help@cornell.edu

🌐 data.research.cornell.edu

Wendy Kozlowski
Data Curation Specialist
RDMSG Coordinator
wak57@cornell.edu

NSF DMPs – a few logistics

- Supplementary document; 2 page limit
 - DMP can roll into project summary pages if needed, but DMP pages cannot be used for project summary
- DMP can simply state that **no data** are to be collected
- Subject to peer review (intellectual merit and/or broader impacts)
- Allowable DM **costs** should be included in budget line G2 and explained in budget justification

<http://www.nsf.gov/bfa/dias/policy/dmp.jsp>

https://www.nsf.gov/pubs/policydocs/pappg19_1/pappg_2.jsp#lIC2j

What NSF (generally) wants to see in a DMP

1. Expected data and research products
2. Formats and standards for data and metadata
3. Policies for access and sharing (dissemination)
 - Timeframe for availability
 - Data storage and security
4. Policies and provisions for re-use, re-distribution and production of derivatives
5. Plan for archiving and preservation of access
 - Period of data retention
6. Roles and responsibilities

NIH Genomic Data Sharing Policy

- Applies to all NIH-funded research that generates large-scale human or non-human genomic data, regardless of the funding level.
- Applies to large-scale data including genome-wide association studies (GWAS*), single nucleotide polymorphisms (SNP) arrays, and genome sequence, transcriptomic, epigenomic, and gene expression data.
- Basic plans must be included in the Resources Sharing Plan section of proposals; detailed plans must be provided prior to award.
- Investigators should contact appropriate IC Program Official or Project Officer to discuss data sharing expectations and timelines.

**GWAS applicants have had a data sharing requirement in place since 2008 (NCBI)*

<http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html#data>
<https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing/>

What NIH (generally) wants to see in a DSP

1. Schedule for sharing
2. Dataset format(s)
3. Description of accompanying data documentation
4. Plan to (or not to) provide analytical tools
5. Plan to (or not to) require data-sharing agreements; when required what that will include (provisions for re-use)
6. Mode of data sharing