Research Data Management Basics

Faculty Development Day
January 24, 2014

Robin Camille Davis
Emerging Technologies & Distance Services Librarian
John Jay Library
Data curation

“The active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities across the sciences, social sciences, and the humanities.”

Source:
- guide.dhcuration.org/faq
- Craigin et al “An Educational Program on Data Curation” (2007)
Research Data Management

Resources and tools for managing your research data. (For faculty and graduate students.)


Welcome!
This guide is intended for John Jay faculty and graduate students who work with research data. You might be new to data management or on old hand — either way, you will find usefull resources here.

How and why do you 'manage data'?
Throughout your research project, you may be handling a lot of data: financial data, crime statistics, text corpora, biological data, etc. It is necessary that you understand how to stay organized and adhere to standards and best practices. Good data management leads to successful projects.

It will also be important for people outside of your project to be able to find and navigate your data in case you must provide further evidence in the form of raw data and if you decide to share the data as supplemental article material or archive it in a repository.

Furthermore, funders like the National Science Foundation (NSF) and the National Institutes of Health (NIH) require data management plans as part of grant applications. It is important for you to know what they expect, since part of your grant’s chance of acceptance depends on a good data management plan.

General data management resources
- About data management + glossary
  The University of Minnesota explains what data management entails and provides a glossary of data management-related terms.
- Guide to managing your data
  A subject guide from MIT that provides information, links, and other resources

Subject Specialist
Robin Davis
Emerging Tech Librarian

Contact Info
(212) 237-8281
Send Email

Links:
- Profile & Guides

Is this guide useful?
Was this information helpful?
☐ ☐ ☐ Don’t
Yes No Know

How useful is this page?
[1 = Not Useful, 5 = Very Useful]
☐ ☐ ☐ ☐ ☐

Additional comments:
What is data?

Tables & databases!

Images! “Narcotics valued at $500,000, seized in hotel raid.”

Text corpora!

Code & software!

Audio & transcripts! http://dc.lib.jjay.cuny.edu/index.php/Detail/Object/Show/object_id/1
Why care about data management?

- Good data management leads to successful projects
- Funders require good data management
- Sharing & reusing data depends on good data management from the get-go
Data management plans

– Summarize your project
– Who is responsible for the data?
– List the kinds of data your project will create and which formats you'll be using
– Identify the standards to which your data will adhere
– Define your plans for
  • data storage and security
  • sharing the data and policies of reuse
  • how you will archive and preserve the data
DATA CURATOR HULK
@DATACURATORHULK

YOU DOWN WITH D-M-P?
YEAH YOU KNOW ME
WHO'S DOWN WITH D-M-P?
EVERY LAST FUNDING AGENCY

8:22 AM - 26 Sep 2013
**Example: Psychology (NSF)**

---

**DATA MANAGEMENT PLAN**

**Introduction**

This Data Management Plan (DMP) covers the data which will be collected under the proposed project entitled: **Causal inference in visual identification: integrating models, experimental paradigms, and behavioral measures** and is designed to be consistent with the NSF Policy on Dissemination and Sharing of Research Results. In accordance with this policy, this plan does not include preliminary analyses (including raw data), drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. Furthermore, data to enable peer review and publication/dissertation analysis to protect intellectual property may be temporarily withheld from distribution and other proposed data management. This plan will make certain that the data produced during the period of this project is appropriately managed to ensure its usability, access, and preservation.

**Description of Data**

The types of data to be produced in the course of the project include background demographic information, accuracy and reaction time for the proposed visual discrimination task, and the results from computational model simulations. The demographic data is needed for research reports to convey the characteristics of the subject population.

**Standards for Data and Metadata Format and Content**

For preservation and long-term access, data collection will be accompanied with proper documentation and associated metadata. This will include the data itself stored as Matlab (.mat) files, documentation files which include a description of how the data was collected, and metadata in the form of the Matlab metafile (Psychtoolbox toolbox) that administers each behavioral task.

**Data Access, Sharing and Archiving**

The researchers associated with this study are not aware of any reasons that might prohibit the sharing of the data to be generated under this project for public use and potential secondary uses. The principal investigator retains the right for final use of the data. The principal investigator will work with their NSF Program Officer to identify public databases which would be appropriate to house data generated from this project as well as for long-term archiving of the materials. Until implementation, public databases are identified, access to the de-identified data will be provided by consulting the PI. For journal publications, summary forms of the data (means and standard deviations) will be published in tables, appendices, or online supplementary materials. Raw data will in principle be available for access and sharing as soon as it is reasonably possible, namely at not longer than one year after submission of the data. The new data from the proposed research will be archived locally by the researchers, on regularly backed-up computers and will be preserved for at least three years beyond the award period, as required by NSF guidelines.

**Data Confidentiality**

Research records will be kept confidential, and access will be limited to the PI and primary research team members. For each testing session, the recorded data will have any identifying information removed and will be released with study code numbers. A database which retains study code numbers to consent forms and identifying information will be stored separately on password-protected computers in a secured, locked office. These computers are housed in research facilities in the Psychology Building at Indiana University-Bloomington, and in the Psychology Department at UCSB. All names at the names of individuals who have participated in each study will be maintained in order to ensure that no individual is tested more than once on related studies. To maintain the privacy of the participants, any record of individual data will only consist of performance measures without any demographic or identifying information.

**Intellectual Property and Sharing of Research Resources**

Intellectual property and data generated under this project will be administered in accordance with both University and NSF policies. Ownership of sole or joint inventions developed under the project will be owned by the institution(s).
This is an example Data Management Plan created by participants of a DataONE best Practices Workshop.

Data Management Plan

I. Types of Data

The products of this research are software to create maps, which are machine- and human-readable, containing information about the format and content of an HDF file. These maps can be used in the future to develop new software that enables reading of HDF files.

Map generation software will be developed for all platforms currently supported. Map generation software will be tested on the entire suite of HDF file types currently held by NASA. For quality control, individual archives that use the software to map their HDF data are responsible for verifying the quality of the maps created by performing a bit-by-bit comparison of data read using the maps and data read using existing HDF reading tools, presumed on a randomly chosen statistically significant subset of each type of HDF file in their holdings. To ensure the self-describing nature of the map files, non-HDF literate, non-native English speaking programming personnel will be used to develop the demonstration software described below. This will verify that the instruction and content of the map files are broadly understandable. The map files created as a result of this project will have the same name as the original HDF file from which they are derived, with the additional extension .map.

One ancillary data product produced by this project will be demonstration software (a reader) that uses the maps to read the HDF files, along with example input/output files and contextual information. While the existence of this demonstration software is not necessary to allow users in the future to read HDF data files, having a demonstration reader available in the near term will help the user community learn how to use the map files. The reader and associated files will be available as open source software at sourceForge.org, along with a copy held at the National Snow and Ice Data Center (NSIDC). Subversion will be used for source code and associated file version control.

Another ancillary product of this proposed work will be a collection of one of each kind of HDF file contained in the NASA archives. This collection of HDF files will be compiled and made available from the NSIDC archive.

The data manager during and after the project for the HDF file collection and the reader software will be Duerr. The HDF Group is responsible for management of the software during the project.

II. Data and Metadata Standards

The maps that will be produced are self-describing and require no documentation. Once created, the maps will become part of the data product, and will be included as one of the files of the HDF archive.
III. Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy

The user community for the map writer is expected to be any repository with HDF data. The user community for the map reader is expected to any user of HDF data in the future who wishes to re-use or adapt map-reading software to their own purposes. The user community for the collection of HDF files is expected to the community of tool developers who need a wide range of data types with which to test their product.

IV. Policies and Provisions for Re-Use, Re-Distribution

This proposed project will create ancillary metadata that will enable future users to read all HDF (Hierarchical Data Format) formatted data. The reader software is open source (licensed under GPLv3) and is freely available. The HDF data at NSIDC is freely available with no restrictions. The collection of HDF files has a suggested citation, which is described in the metadata.

V. Plans for Archiving and Preservation of Access

The writer software will be preserved by the HDF Group for the life of the HDF libraries. The HDF Group uses industry-standard best practices to ensure the integrity of their software and systems. Once the map writer has been used to generate maps for every HDF file in existence, the continued existence of the writer software is not required. The reader software will be preserved at SourceForge.org for as long as there is community interest. The collection of HDF files will be preserved at NSIDC as long as utility is deemed high.
NSF-EHR: Education and Human Resources: 2. Period of data retention

EHR is committed to timely and rapid data distribution. However, it recognizes that types of data can vary widely and that acceptable norms also vary by scientific discipline. It is strongly committed, however, to the underlying principle of timely access, and applicants should address how this will be met in their Data Management Plan.

Progress

Click on a section below to edit it at any time.

- = complete

Plan description

1. Data generated by the project
2. Period of data retention
3. Data format and dissemination
4. Data storage and preservation of access
5. Additional possible data management requirements

Help

This section will allow you to account for any delay in the accessibility of your data after your research is done. Consider any reasons why you would not make the data immediately available — for instance, maybe you have political, commercial, or patent concerns that will require you to postpone access to the data you produce. Consider these questions:

- How long will the original data collector/creator/principal investigator retain the right to use the data before opening it up to wider use?

Resources

General
- NSF Data Sharing Policy
- NSF Data Management Plan Requirements
- ICPSR – Effective Data Management
- NSF EHR Directorate-wide Guidance
- DataCONE Best Practices: Sharing Data: Legal and Policy Considerations
Metadata!

• “data about data”
• constructed information
• developed by people for a purpose

Source:
Metadata

• Disciplinary standards
  – Ecological Metadata Language
  – International virtual observatory specifications
  – Text Encoding Initiative

• Discovery standards
  – Title? Author? Date?
Data storage
(during research)

• Lots Of Copies Keep Stuff Safe (LOCKSS)

• Automate & test your backup system
Data security

A GUIDE TO PROTECTING YOUR COMPUTER & PERSONAL INFORMATION

WHAT YOU NEED TO KNOW?

The Department of Information Technology (DoIT) is responsible for safeguarding John Jay computers and networks from security breaches. We at DoIT are proactive in protecting the college's computing assets with regular systems upgrades, software patches, anti-virus software, etc. However, if you install spurious software or applications such as games, music or videos either from an untrusted source on the Internet or from infected media such as a flash drive or CD your computer can get infected and become compromised.

Hackers who gain control of on campus computers then use them to further attack other computers or the entire network system of the College. In most cases the focus of these hackers is to extract confidential data they can use for financial gain. It is important that as a technology user at John Jay you understand the responsibility you share with DoIT to protect against this ever present threat.

SECURITY STARTS WITH YOU! DO YOUR PART!

The most important component of good computer security for the John Jay Community is not a firewall or some network security device but your understanding of what risks exist and what you can do to safeguard yourself and others against them.

AWARENESS IS YOUR BEST DEFENSE! EDUCATE YOURSELF AND EDUCATE OTHERS!
Data security

• CUNY requires encryption of all confidential data
• Store non-public university data on a secure server (not your computer or flash drive)
• CUNY policies: security.cuny.edu
Archiving your data

3 years? 5 years? Forever?

Long-term storage:
• provide documentation
• prepare the data for long-term storage
• straighten out institutional/organizational support
• archiving solutions:
  – department or institution
  – trusted repository
Data repositories

- online vault for materials contributed by many people.
- its contents can be publicly available, restricted to certain people, or embargoed
Find Data

More Search Options
- Browse by topic
- Browse by series
- Browse by geography
- Browse by investigator
- Browse international data
- View all studies
  - View all studies for which online analysis is available
  - View all replication datasets (more info)
  - View all studies that have learning guides

Restricted Data

ICPSR ensures respondent confidentiality within public-use versions of the datasets. The vast majority of ICPSR data holdings are public-use files with no restrictions on their access. Sometimes the protective measures taken to reduce disclosure risk would significantly degrade the research potential of the data. In these cases, ICPSR provides access to restricted use versions that retain confidential data by imposing stringent requirements for accessing them. (more)
Sharing your data

• Reproducibility
• Funder requirements (e.g. NIH)
  – White House memo: research funded by major federal agencies must be publicly accessible
• For science! / For the humanities!
• Impact
“Principal Findings:
• We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. ...
• Publicly available data was significantly (p = 0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.”

Source:
Thanks! Questions?

Robin Camille Davis
Emerging Technologies & Distance Services Librarian
John Jay Library

robdavis@jjay.cuny.edu « not rdamis!

@robinacamille

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.