Data Management General Guidance

Table of Contents

Introduction

What is a 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. Many data management issues can be handled easily or avoided entirely by planning ahead. With the right process and framework it doesn't take too long and can pay-off enormously in the long run.

Who requires a plan?

In February of 2013, the White House Office of Science and Technology Policy (OSTP) (http://www.whitehouse.gov/administration/eop/ostp) issued a memorandum (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf) directing Federal agencies that provide significant research funding to develop a plan to expand public access to research. Among other requriments, the plans must

> Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified

The National Science Foundation (NSF) already requires a 2-page plan as part of the funding proposal process. Soon most or all US Federally funded grants will require some form of data management plan.

We can help

We have been working with internal and external partners to make data management plan development less complicated. By getting to know your research and data, we can match your specific needs with data management best practices in your field to develop a data management plan that works for you. If you do this work at the beginning of your research process, you will have a far easier time following-through and complying with funding agency and publisher requirements.

We recommend that those applying for funding from US Federal agencies, such as the NSF, use the DMPTool. The DMPTool provides guidance for many of the NSF Directorate and Division requirements, along with links to UC resources, services, and help. Contact UC3 (http://www.cdlib.org/services/uc3/contact.html) if you have questions and would like feedback on your plan.

Types of Data

Research projects generate and collect countelss varieties of data. To forumulate a data management plan, it's useful to categorize your data in four ways: by source, format, stability, and volume.

What's the source of the data?

Although data comes from many different sources, but they can be grouped into four main categories. The category(ies) your data comes from will affect the choices that you make throughout your data management plan.

Observational

- Captured in real-time, typically outside the lab
- Usually irreplaceable and therefore the most important to safeguard
- Examples: Sensor readings, telemetry, survey results, images

Experimental

- Typically generated in the lab or under controlled conditions
- Often reproducible, but can be expensive or time-consuming
- Examples: gene sequences, chromatograms, magnetic field readings

Simulation

- Machine generated from test models
- Likely to be reproducible if the model and inputs are preserved
- Examples: climate models, economic models

Derived / Compiled

- Generated from existing datasets
- Reproducible, but can be very expensive and time-consuming
- Examples: text and data mining, compiled database, 3D models

What's the form of the data?

Data can come in many forms, including

- **Text:** field or laboratory notes, survey responses
- **Numeric:** tables, counts, measurements
- **Audiovisual:** images, sound recordings, video
- **Models, computer code**
- **Discipline-specific:** FITS in astronomy, CIF in chemistry
- **Instrument-specific:** equipment outputs

How stable is the data?

Data can also be fixed or changing over the course of the project (and perhaps beyond the project's end). Do the data ever change? Do they grow? Is previously recorded data subject to correction? Will you need to keep track of data versions? With respect to time, the common categories of dataset are

- **Fixed datasets:** never change after being collected or generated
- **Growing datasets:** new data may be added, but the old data is never changed or deleted
- **Revisable datasets:** new data may be added, and old data may be changed or deleted

The answer to this question affects how you organize the data as well as the level of versioning you will need to undertake. Keeping track of rapidly changing datasets can be a challenge, so it is imperative that you begin with a plan to carry you through the entire data management process.

How much data will the project produce?

For instance, image data typically requires a lot of storage space, so you'll want to decide whether to retain all your images (and, if not, how you will decide which to discard) and where such large data can be housed. Be sure to know your archiving organization's capacity for storage and backups.

To avoid being under-prepared, estimate the growth rate of your data. Some questions to consider are

- Are you manually collecting and recording data?
- Are you using observational instruments and computers to collect data?
- Is your data collection highly iterative?
- How much data will you accumluate every month or every 90 days?
- How much data do you anticipate collecting and generating by the end of your project?

File Formats

The file format you choose for your data is a primary factor in someone else's ability to access it in the future. Tthink carefully about what file format will be best to manage, share, and preserve your data. Technology continually changes and all contemporary hardware and software should be expected to become obsolete. Consider how your data will be read if the software used to produce it becomes unavailable. Although any file format you choose today may become unreadable in the future, some formats are more likely to be readable than others.

Formats likely to be accessible in the future are:

- Non-proprietary
- Open, with documented standards
- In common usage by the research community
- Using standard character encodings (i.e., ASCII, UTF-8)

- Uncompressed (space permitting)

Examples of preferred format choices:

- Image: JPEG, JPG-2000, PNG, TIFF
- Text: plain text (TXT), HTML, XML, PDF/A
- Audio: AIFF, WAVE
- Containers: TAR, GZIP, ZIP
- Databases: prefer XML or CSV to native binary formats

Examples of discouraged format choices and better alternatives:

Discouraged Format	**Alternative Format**
Excel (.xls, .xlsx)	Comma Separated Values (.csv)
Word (.doc, .docx)	plain text (.txt), or if formatting is needed, PDF/A (.pdf)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf)
Photoshop (.psd)	TIFF (.tif, .tiff)
Quicktime (.mov)	MPEG-4 (.mp4)

If you find it necessary or convenient to work with data in a proprietary/discouraged file format, do so, but consider saving your work in a more archival format when you are finished.

For more information on recommended formats, see the CDL Digital File Format Recommendations (http://www.cdlib.org/gateways/docs/cdl_dffr.pdf).

Tabular data

Tabular data warrants special mention because it is so common across disciplines, mostly as Excel spreadsheets. If you do your analysis in Excel, you should use the "Save As..." command to export your work to .csv format when you are done. Your spreadsheets will be easier to understand and to export if you follow best practices when you set them up, such as:

- Don't put more than one table on a worksheet
- Include a header row with understandable title for each column
- Create charts on new sheets- don't embed them in the worksheet with the data

Other risks to accessibility

- Encrypted data may be effectively lost if it was encrypted with a key that has been lost (e.g., a forgotten password). For this reason, encrypted data representations are strongly discouraged.
- Data that is legally encumbered may also be considered lost. So may data bound by ambiguous or unknown access and archiving rights, because the cost of clarifying the rights situation is often prohibitive. See data rights and licensing for guidance.

Organizing Files

Basic Directory and File Naming Conventions

These are rough guidelines to follow to help manage your data files in case you don't already have your own internal conventions. When organizing files, the top-level directory/folder should include:

- Project title
- Unique identifier (Guidance on persistent external identifiers is available)
- Date (yyyy or yyyy.mm.dd)

The sub-directory structure should have clear, documented naming conventions. Separate files or directories could apply, for example, to each run of an experiment, each version of a dataset, and/or each person in the group.

- Reserve the 3-letter file extension for the file format, such as .txt, .pdf, or .csv.
- Identify the activity or project in the file name.
- Identify separate versions of files and datasets using file or directory naming conventions. It can quickly become difficult to identify the 'correct' version of a file.
- Record all changes to a file no matter how small. Discard obsolete versions after making backups.

File Renaming

Tools to help you:

- Bulk Rename Utility (http://www.bulkrenameutility.co.uk) (Windows; free)

- Renamer (http://renamer.com) (Mac; free trial)
- PSRenamer (http://www.powersurgepub.com/products/psrenamer.html) (Linux, Mac, Windows; free)

File Naming Conventions for Specific Disciplines

Many disciplines have recommendations, for example:

- DOE's Atmospheric Radiation Measurement (ARM) program (http://www.arm.gov/data/docs/plan)

Metadata: Data Documentation

Why document data?

Clear and detailed documentation is essential for data to be understood, interpreted, and used. Data documentation describes the content, formats, and internal relationships of your data in detail and will enable other researchers to find, use and properly cite your data.

Begin to document your data at the very beginning of your research project and continue throughout the project. Doing so will make the process much easier. If you have to construct the documentation at the end of the project, the process will be painful and important details will have been lost or forgotten. Don't wait to document your data!

What to document?

Research Project Documentation

- Rationale and context for data collection
- Data collection methods
- Structure and organization of data files
- Data sources used (see citing data)
- Data validation and quality assurance
- Transformations of data from the raw data through analysis
- Information on confidentiality, access & use conditions

Dataset documentation

- Variable names and descriptions
- Explanation of codes and classification schemes used
- Algorithms used to transform data (may include computer code)
- File format and software (including version) used

How will you document your data?

Data documentation is commonly called metadata – "data about data". Researchers can document their data according to various metadata standards. Some metadata standards are designed for the purpose of documenting the contents of files, others for documenting the technical characteristics of files, and yet others for expressing relationships between files within a set of data. If you want to be able to share or publish your data, the DataCite metadata standard (http://schema.datacite.org/) is of particular signficiance. It is important to establish a metadata strategy that is capable of describing your data and satisfying your data management needs. For assistance in defining an adequate metadata strategy, please contact uc3@ucop.edu (mailto:uc3@ucop.edu).

Below are some general aspects of your data that you should document, regardless of your discipline. At minimum, store this documentation in a "readme.txt" file, or the equivalent, with the data itself. You can also reference a published article that may contain some of this information.

General Overview	
Title	Name of the dataset or research project that produced it
Creator	Names and addresses of the organizations or people who created the data; preferred format for personal names is surname first (e.g., Smith, Jane).
Identifier	Unique number used to identify the data, even if it is just an internal project reference number
Date	Key dates associated with the data, including: project start and end date; release date; time period covered by the data; and other dates associated with the data lifespan, such as maintenance cycle, update schedule; preferred format is yyyy-mm-dd, or yyyy.mm.dd-yyyy.mm.dd for a range
Method	How the data were generated, listing equipment and software used (including model and version numbers), formulae, algorithms, experimental protocols, and other things one might include in a lab notebook
Processing	How the data have been altered or processed (e.g., normalized)
Source	Citations to data derived from other sources, including details of where the source data is held and how it was accessed
Funder	Organizations or agencies who funded the research

Content Description	

Subject	Keywords or phrases describing the subject or content of the data
Place	All applicable physical locations
Language	All languages used in the dataset
Variable list	All variables in the data files, where applicable
Code list	Explanation of codes or abbreviations used in either the file names or the variables in the data files (e.g. '999 indicates a missing value in the data')

Technical Description	
File inventory	All files associated with the project, including extensions (e.g. 'NWPalaceTR.WRL', 'stone.mov')
File Formats	Formats of the data, e.g., FITS, SPSS, HTML, JPEG, etc.
File structure	Organization of the data file(s) and layout of the variables, where applicable
Version	Unique date/time stamp and identifier for each version
Checksum	A digest value computed for each file that can be used to detect changes; if a recomputed digest differs from the stored digest, the file must have changed
Necessary software	Names of any special-purpose software packages required to create, view, analyze, or otherwise use the data

Access	
Rights	Any known intellectual property rights, statutory rights, licenses, or restrictions on use of the data
Access information	Where and how your data can be accessed by other researchers

Persistent Identifiers

If you want to be able to share or cite your dataset, you'll want to assign a public persistent unqiue identifier to it. There are a variety of public identifier schemes, but common properties of good schemes are that they are:

- Actionable (you can "click" on them in a web browser)
- Globally unique across the internet
- Persistent for at least the life of your data

Today, this means data identifiers should fit inside a web address (URL) and be well-enough managed to remain actionable over the long-term. The most important factors in long-term data sharing are stable data storage and well-managed identifier redirection.

Any URL can be thought of as "resolving" either directly to its target (via the URL's hostname) or indirectly through one or more "redirects" to a final target URL. If your dataset is moved (e.g., from one archive to another) redirection allows the identifier to continue to resolve to the dataset, now at its new location.

An important factor to consider is choice of URL hostname. This is the domain name at the beginning of a URL (right after the "http://") that determines where URL resolution starts (for example, daac.ornl.gov). An identifier that doesn't contain a hostname may implicitly use a well-known hostname as the starting point for resolution. For example, dx.doi.org is the hostname for DOIs, so the document identified by doi:10.1000/182 can be found by typing "http://dx.doi.org/10.1000/182" in a web browser.

Because persistent (long-term) identifiers tend to be opaque (e.g., a string of digits) and reveal little or nothing about the nature of the identified object, it is also important for you to maintain metadata associated with the object. Among the most important pieces of metadata for you to maintain is the target URL that ensures that the identifier remains actionable. Whatever identifier scheme you choose, if you don't update the target URL when your data is moved, the identifier will break.

Here are some identifier schemes:

- ARK (http://www.cdlib.org/services/uc3/curation/ark.html) (Archival Resource Key) – a URL with extra features allowing you to ask for descriptive and archival metadata and to recognize certain kinds of relationships between identifiers. ARKs are used by memory organizations such as libraries, archives, and museums. They are resolved at "http://www.nt2.net". Resolution depends on HTTP redirection and can be managed through an API or a user interface. There are no usage fees.
- DOI (http://www.doi.org/) (Digital Object Identifier) – an identifier that becomes actionable when embedded in a URL. DOIs are very popular in academic journal publishing. They are resolved at "http://dx.doi.org". Resolution depends on HTTP redirection and the Handle identifier protocol, and can be managed through an API or a user interface. Annual fees apply to each DOI.
- Handle (http://www.handle.net/) – an identifier that becomes actionable when embedded in a URL. Handles are resolved at "http://handle.net". Resolution depends on HTTP redirection and the Handle protocol, and can be managed through an API or a user interface. Annual fees apply to each local Handle server.
- InChI (http://www.iupac.org/inchi/) (IUPAC International Chemical Identifier) – a non-actionable identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of diverse data compilations.
- LSID (http://en.wikipedia.org/wiki/LSID) (Life Sciences Identifier) – a kind of URN that identifies a biologically significant resources, including

species names, concepts, occurrences, and genes or proteins, or data objects that encode information about them. Like other URNs, it becomes actionable when embedded in a URL.

- NCBI (http://www.ncbi.nlm.nih.gov/Sequin/acc.html) (National Center for Biotechnology Information) ACCESSION – a non-actionable number in use by NCBI.
- PURL (http://www.purl.org/) (Persistent Uniform Resource Locator) – a URL that is always redirected through a hostname (often purl.org). Resolution depends on HTTP redirection and can be managed through an API or a user interface. There are no usage fees.
- URL (Uniform Resource Locator) – the typical "address" of web content. It is a kind of URI (Uniform Resource Identifier) that begins with "http://" and consists of a string of characters used to identify or name a resource on the Internet. Such identification enables interaction with representations of the resource over a network, typically the World Wide Web, using the HTTP protocol. Well-managed URL redirection can make URLs as persistent as any identifier. Resolution depends on HTTP redirection and can be managed through an API or a user interface. There are no usage fees.
- URN (Uniform Resource Name) – an identifier that becomes actionable when embedded in a URL. Resolution depends on HTTP redirection and the DDDS protocol, and can be managed through an API or a user interface. A browser plug-in can save you from typing a hostname in front of it. There are no usage fees.

EZID: Identifiers Made Easy

CDL provides an identifier service called EZID (http://www.cdlib.org/services/uc3/ezid/) that offers several choices of identifier. EZID enables you to take control of the management and distribution of your datasets, share and get credit for your datasets, and build your reputation for the collection and documentation of research. By making data resources easier to access, re-use, and verify, EZID helps you to build on previous work, conduct new research, and avoid duplicating previous work.

Security and Storage

Data Security

Data security is the protection of data from unauthorized access, use, change, disclosure and destruction. Make sure your data is safe in regards to:

- Network security
  - Keep confidential data off the Internet
  - In extreme cases, put sensitive materials on computers not connected to the internet
- Physical security
  - Restrict access to buildings and rooms where computers or media are kept
  - Only let trusted individuals troubleshoot computer problems
- Computer systems and files
  - Keep virus protection up to date
  - Don't send confidential data via e-mail or FTP (or, if you must, use encryption)
  - Set passwords on files and computers
  - React with skepticism to phone calls and emails that claim to be from your insitution's IT department

Encryption and Compression

Unencrypted data will be more easily read by you and others in the future, but you may need to encrypt sensitive data.

- Use mainstream encryption tools (e.g., PGP)
- Don't rely on 3rd party encryption alone
- Keep passwords and keys on paper (2 copies)

Uncompressed data will be also be easier to read in the future, but you may need to compress files to conserve disk space.

- Use a mainstream compression tool (e.g., ZIP, GZIP, TAR)
- Limit compression to the 3rd backup copy

Backups and storage

Making regular backups is an integral part of data management. You can backup data to your personal computer, external hard drives, or departmental or university servers. Software that makes backups for you automatically can simplify this process considerably. CDs or DVDs are not recommended because they are easily lost, decay rapidly, and fail frequently. The UK Data Archive provides additional guidelines on data storage, backup and security (http://www.data-archive.ac.uk/create-manage/storage).

Backup Your Data

- Good practice is to have three copies in at least two locations (e.g. original + external/local backup + external/remote backup)
- Geographically distribute your local and remote copies to reduce risk of calamity at the same location (power outage, flood, fire, etc.)

Data Backup Options

- Hard drive using software like:
  - Windows 8 File History (http://windows.microsoft.com/en-us/windows-8/set-drive-file-history)
  - OS X Time Machine (http://www.apple.com/support/timemachine/)
  - Linux/UNIX rsync (http://en.wikipedia.org/wiki/Rsync))
- Tape backup system
- Many institutions provide a service similar to UCBackup (http://ist.berkeley.edu/ucbackup/) at UC Berkeley. Check with your campus IT support to see if backup service is available. Althernately, your academic department may provide storage space and backup services.
- Cloud storage - some examples of private sector storage resources include:
  - Amazon S3 (http://aws.amazon.com/s3/) and Glacier (http://aws.amazon.com/glacier/)—Requires client software, no encryption support
  - S3-based Remote Hard Drive Services such as Elephant Drive (http://home.elephantdrive.com/) and Jungle Disk (https://www.jungledisk.com/).
  - Mozy (http://mozy.com/) (from EMC) Free client software, 448-bit Blowfish encryption or AES key
  - Carbonite Free (http://www.carbonite.com/)

Test your backup system

To be sure that your backup system is working, periodically retrieve your data files and confirm that you can read them. You should do this when you initially set up the system and on a regular schedule thereafter.

Other data preservation considerations

Who is responsible for managing and controlling the data?

Who controls the data (e.g., the PI, a student, your lab, your university, your funder)? Before you spend a lot of time figuring out how to store the data, to share it, to name it, etc. you should make sure you have the authority to do so.

For what or whom are the data intended?

Who is your intended audience for the data? How do you expect they will use the data? The answer to these questions will help inform structuring and distributing the data.

How long should the data be retained?

Is there any requirement that the data be retained? If so, for how long? 3-5 years, 10-20 years, permanently? Not all data need to be retained, and some data required to be retained need not be retained indefinitely. Have a good understanding of your obligation for the data's retention.

Beyond any externally imposed requirments, think about the long-term usefulness of the data. If the data is from an experiment that you anticipate will be repeatable more quickly, inexpensively, and accurately as technology progresses, you may want to store it for a relatively brief period. If the data consists of observations made outside the laborartory that can never be repeated, you may wish to store it indefinitely.

Sharing and Archiving

Why share your data?

- Required by publishers (e.g., Cell (http://www.cell.com/authors#ed_policies), Nature (http://www.nature.com/authors/policies/availability.html), Science (http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail)).
- Required by government funding agencies (e.g., NIH (http://grants.nih.gov/grants/policy/data_sharing/), NSF (http://www.nsf.gov/bfa/dias/policy/dmp.jsp))
- Allows data to be used to answer new questions
- Makes research more open
- Makes your papers more useful and citable by other researchers

Considerations when preparing to share data

- **File Formats for Long Term Access:** The file format in which you keep your data is a primary factor in one's ability to use your data in the future. Plan for both hardware and software obsolescence. See file formats and organization for details on long-term storage formats.
- **Don't Forget the Documentation:** Document your research and data so others can interpret the data. Begin to document your data at the very beginning of your research project and continue throughout the project. See data documentation and metadata for details.
- **Ownership and Privacy:** Make sure that you have considered the implications of sharing data in terms of copyright, IP ownership, and subject confidentiality. See copyright and confidentiality for details.

Ways to share your data

- Email to individual requesters
- Post online via a project or personal web site
- Submit as supplemental material to be hosted on a journal publisher's website

- Deposit in an open repository or archive
- Deposit in an open repository and publish a "data paper" describing the data

While the first three options above are valid ways to share data, a repository is much more able to provide long-term access. Data deposited in a repository can be supplemented with a "data paper"—a relatively new type of publication that describes a dataset, but does not analyze it or draw any conclusions—published in a journal such as Nature Scientific Data (http://www.nature.com/scientificdata/) or [Geoscience Data Journal] (http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060).

Finding a data repository

You should select a repository or archive for your data based on the long-term security offered and the ease of discovery and access by colleagues in your field. There are two common types of repository to look for:

- **Discipline specific:** accepts data in a particluar field or of a particluar type (e.g., GenBank (http://www.ncbi.nlm.nih.gov/genbank/) accepts nucleotide sequence data)
- **Institutional:** accepts data of any type proudced within the institution that maintains it (e.g., the University of California's Merritt (http://www.cdlib.org/services/uc3/merritt/))

A searchable and browsable list of repositories can be found at these websites:

- DataBib (http://databib.org/): a directory of research data repositories
- re3data.org (http://www.re3data.org): a REgistry of REsearch data REpositories
- Data Repositories (http://oad.simmons.edu/oadwiki/Data_repositories) in the Open Access Directory: a list of repositories hosted by Simmons College
- BioSharing (http://www.biosharing.org/): a directory of life sciences databases and reporting standards

Citing Data

Citing data is important in order to:

- Give the data producer appropriate credit
- Allow easier access to the data for re-purposing or re-use
- Enable readers to verify your results

Citation Elements

A dataset should be cited formally in an article's reference list, not just informally in the text. Many data repositories and publishers provide explict instructions for citing their contents. If no citation information is provided, you can still construct a citation following generally agreed-upon guidelines from sources such as the CODATA Report on data citation (https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article), the ESIP Federation Data Citation Guidelines (http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations), and the current DataCite Metadata Schema (http://schema.datacite.org/).

Core elements

There are 5 core elements usually included in a dataset citation, with additional elements added as appropriate.

- **Creator(s)** – may be individuals or organizations
- **Title**
- **Publication year** when the dataset was released (may be different from the **Access date**)
- **Publisher** – the data center, archive, or repository
- **Identifier** – a unique public identifier (e.g., an ARK or DOI)

Creator names in non-Roman scripts should be transliterated using the ALA-LC Romanization Tables (http://www.loc.gov/catdir/cpso/roman.html).

Common additional elements

Although the core elements are sufficient in the simplest case – citation to the entirety of a static dataset – additional elements may be needed if you wish to cite a dynamic dataset or a subset of a larger dataset.

- **Version** of the dataset analyzed in the citing paper
- **Access date** when the data was accessed for analysis in the citing paper
- **Subset** of the dataset analyzed (e.g., a range of dates or record numbers, a list of variables)
- **Verifier** that the dataset or subset accessed by a reader is identical to the one analyzed by the author (e.g., a Checksum (http://en.wikipedia.org/wiki/Checksum))
- **Location** of the dataset on the internet, needed if the identifier is not "actionable" (convertable to a web address)

Example citations

- Laurance, Megan (2013): Re-analysis of microarray data from rapamycin resistant DLBCL cell lines. University of California, San Francisco. Dataset.

- Kumar, Sujai (2012): 20 Nematode Proteomes. Figshare. Dataset. doi:10.6084/m9.figshare.96035.
- Morran LT, Parrish II RC, Gelarden IA, Lively CM (2012) Data from: Temporal dynamics of outcrossing and host mortality rates in host-pathogen experimental coevolution. Evolution doi:10.5061/dryad.c3gh6
- Donna Strahan. "Petra Great Temple Excavations: 08-B-1 (Small Find)" (Released 2009-10-26). Martha Sharp Joukowsky (Ed.) Open Context. (http://opencontext.org/subjects/30C3F340-5D14-497A-B9D0-7A0DA2C019F1)
- OECD (2008), Social Expenditures aggregates, OECD Social Expenditure Statistics (database). doi: 10.1787/000530172303 (http://dx.doi.org/10.1787/000530172303) (Accessed on 2008-12-02).
- Cavalieri, D., C. Parkinson, P. Gloersen, and H. J. Zwally. 1996, updated 2006. Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data, March 2002 – Sept. 2003. Boulder, Colorado USA: National Snow and Ice Data Center. (http://nsidc.org/data/nsidc-0051.html) (Accessed on 2008-05-14).
- Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. (http://dx.doi.org/10.1594/WDCC/dphase_mpeps)
- Manoug, J L (1882): Useful data on the rise of the Nile. Alexandria : Printing-Office V Penasson. (http://n2t.net/ark:/13960/t44q88124)

Copyright and Privacy

Sharing data that you produced/collected yourself

- **Data is not copyrightable.** However, a presentation of data (such as a chart or table) may be.
- **Data can be licensed.** Some data providers apply licenses that limit how the data can be used to protect the privacy of study participants or to guide downstream uses of the data (e.g., requiring attribution or forbidding for-profit use)
- If you want to promote sharing and unlimited use of your data, you can make your data available under a Creative Commons CC0 Declaration (http://creativecommons.org/choose/zero/) to make your wishes explicit.

Sharing data that you have collected from other sources

- You may or may not have the rights to do so, depending upon whether that data were accessed under a license with terms of use.
- Most databases to which the UC Libraries subscribe are licensed and prohibit redistribution of data outside of UC. For more information on terms of use for databases licensed by the Libraries, contact UC3 (http://www.cdlib.org/services/uc3/contact.html).

If you are uncertain as to your rights to disseminate data, UC researchers can consult with your campus Office of General Council. Note: Laws about data vary outside the U.S.

For a general discussion about publishing your data, applicable to many disciplines, see the ICPSR Guide to Social Science Data Preparation and Archiving (http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf).

Confidendiality and Ethical Concerns

It is vital to maintain the confidentiality of research subjects both as an ethical matter and to ensure continuing participation in research. Researchers need to understand and manage tensions between confidentiality requirements and the potential benefits of archiving and publishing the data.

- **Evaluate the anonymity of your data.** Consider to what extent your data contains direct or indirect identifiers (http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/) that could be combined with other public information to identify research participants.
- **Obtain a confidentiality review.** A benefit of depositing your data with ICPSR (http://www.icpsr.umich.edu/) is that their staff offers a Disclosure review service (http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/index.html) to check your data for confidential information.
- **Comply with UC regulations.** researchers concerned about confidentiality issues with their data should consult the UC policy for Protection of Human Subjects in Research (http://policy.ucop.edu/doc/2500499).
- **Comply with regulations for health research** set forth in the Heatlth Insurance Portability and Accountability Act (HIPPA) (http://privacyruleandresearch.nih.gov/).

To ethically share confidential data, you may be able to

- **Gain informed consent** for data sharing (e.g. deposit in a repository or archive)
- **Anonymize** the data by removing identifying information. Be aware, however, that any dataset that contains enough information to be useful will always present some risk.
- **Restrict the use** of your data. The ICPSR DSDR (http://www.icpsr.umich.edu/icpsrweb/DSDR/) provides a tool for Designing a Restricted Data Use Contract (http://www.icpsr.umich.edu/icpsrweb/DSDR/rduc/).