Data Management

By Kristin Briney – University of Wisconsin-Milwaukee
Ye Li – University of Michigan-Ann Arbor
Leah R. McEwen – Cornell University

Learning Objectives

Develop awareness, through examples and lab exercises, of the challenges associated with managing scientific experimental data. Learn best practices for managing data across the entire research lifecycle, from conceiving of an experiment to sharing resulting data and analysis. Students will develop skills in the following areas:

- Know how to organize data to better find it later
- Follow best practices for storage and backup to protect files from loss
- Document data so that anyone can follow what you did
- Properly cite and follow license conditions when using another researcher's data

3.1 Overview

This section of Module 1 covers many small practices that are not regularly taught but can help you manage your scientific and digital information better. Many, many scientists waste time trying to find a certain file in a collection, mentally reconstruct an experiment from insufficient notes, or reproduce whole experiments after accidental data loss. Don't let that be you!

The point of good data management, as the following practices are collectively called, is to make sure you can find and understand your information when you need it, even if that is 10 years after you created the files. Your information has value and this means that you should take care of it properly. Manage your data today and make your life and others' lives much easier tomorrow and down the road.

This section covers a wide range of practices that will help you take care of both your scientific data and your personal files. Note that you do not need to adopt every practice outlined in this section. Rather, try out anything that looks helpful, play around with the recommendations, and figure out good habits that work for you. Alternatively, try adopting one practice at a time until it becomes routine, then repeat with a new practice. Anything you do to take care of your data is better is an improvement and means your data are that much safer.

There are many good things you can do to protect your data from loss, make your files easier to find and use when you need them, or allow a collaborator to easily use your data. These practices occur in a range of categories, which are covered in this Section 3 of Module 1. The categories and practices are as follows:

- File organization
  - File organization
  - File naming

    - A standard for dates
- Storage and backups
    - 3-2-1 Rule
    - Backups
- Data quality control
    - Spreadsheet best practices
- Documentation
    - Taking better notes
    - Data dictionaries
    - Templates
    - README.txt files
- Data sharing
    - Data citation
    - Managing Re-Use Permissions with Creative Commons

3.2 File Organization

File Organization

Organization helps you find and sort through your data and makes it easier to use your data in the future. Save yourself time in the future by making sure that your files are organized as you create them.

The most important thing for organization is to have a system and use it consistently. This will help you track down files when you need them and not waste time combing through useless information.

Here are several options for organization:

- By project
- By analysis type
- By date
- By researcher
- By thesis chapter
- By site or data source

You can also use these systems in combination.

Basically, figure out a system that works for your data (does not have to be listed here) and stick to it. You should also document your organization system in your lab notebook or another prominent place.

Here are several examples of file organization systems:

Experimental data:

- By experiment
  - By file type (raw data, analyzed data, figures, etc.)

Collaborative project data:

- By researcher
  - By project
    - By date

Adapted from the data management guide by UWM Libraries (http://guides.library.uwm.edu/data), CC-BY.

File Naming

HORROR STORIES:

- http://retractionwatch.com/2014/01/07/doing-the-right-thing-authors-retract-brain-paper-with-systematic-human-error-in-coding/

- http://retractionwatch.com/2015/04/29/chem-paper-fails-to-catalyze-when-wrong-files-are-inadvertently-used/

A file naming convention add standardization to your files, making them much easier to organize and locate. It will also help your colleagues sort through your files should you fall ill or leave the lab. Your naming scheme should be documented in your laboratory notebook (preferably at the front or back for easy access) or in a prominent place for this reason.

There are conventions available for you to choose from, though you will probably want to customize one for your own purposes. There are a few general tips for creating systems for naming files.

First, pick a group of files that you wish to name consistently and decide on the key information that will distinguish one file from another. Pick 2-3 things that will tell you a file's contents. Examples are:

- Date
- Site
- Analysis
- Sample
- Short description

Once you pick your key pieces of information, arrange them into a pattern using the following rules:

- Files should be named consistently
- Files names should be descriptive but short (<25 characters)
- Use underscores instead of spaces
- Avoid these characters: " / \ : * ? ' < > [ ] & $

You can also add version information, as necessary. Versioning can be imminently helpful when you are analyzing data. If you make a change to your data that you don't want to keep, it's simple to go back to an earlier version of the file. The same is true if a file gets corrupted or if you simply want to change your analysis method. The key to making versioning work is being consistent with version names, periodically saving to new versions, and documenting the differences between versions.

- For analyzed data, use version numbers
- Save files often to a new version
- Label the final version FINAL

Using these guidelines, here are some example naming conventions and example file names. The first example, in particular, is useful for organizing .pdf's of journal articles.

- AuthorLastName-Year-Title
    - Smith-2010-ImpactOfStressOnSeaMonkeys
    - Hailey-1999-VeryImportantDNAStudy
- YYYYMMDD_site_sampleNumber
    - 20140422_PikeLake_03
    - 20140424_EastLake_12

- Experiment_Analysis_Version
    - KMnO4_FirstOrder_v2
    - HCl_ZeroOrder_v5

Adapted from "Starting Small: File Naming Conventions" by Kristin Briney (http://dataabinitio.com/?p=14/), CC-BY, and the data management guide by UWM Libraries (http://guides.library.uwm.edu/data), CC-BY.

Dates

The standard ISO 8601 is incredibly useful for data management. This standard concerns dates, a common type of information used for data and documentation. To understand why this standard is important, consider the following dates:

- March 5, 2014
- 2014-03-05
- 3/5/14
- 05/03/2014
- 5 Mar 2014

All of these represent the same date but are expressed in different formats. The problem is that if someone uses all of these formats in her notes, how will you ever find everything that happened on March 5th? It's simply too much work to search for all the possible variations. The answer to this problem is ISO 8601.

ISO 8601 dictates that all dates should use the format "YYYYMMDD" or "YYYY-MM-DD". So the example above becomes "20140305" or "2014-03-05". This provides you with a consistent format for all of your dates. Such consistency allows you to more easily find and organize your data, the hallmark of good data management.

ISO 8601's consistency is great but is particularly useful when you use it at the beginning of file names. This is because dates using this standard sort chronologically by year, by month, and then by date. So if you date all of your file names using ISO 8601, you suddenly have a super easy way to find and sort through information.

Adapted from "Dating Your Data (or How I Learned to Stop Worrying and Love the Standard)" by Kristin Briney (http://dataabinitio.com/?p=449), CC-BY.

3.3 Storage and Backup

3-2-1 Rule

VIDEO: https://www.youtube.com/watch?v=_F_r56dkq2I

HORROR STORIES:

- www.flickr.com/photos/gailst/7824341752/
- https://projects.ac/blog/the-stuff-of-nightmares-imagine-losing-all-your-research-data/
- http://chronicle.com/blogs/wiredcampus/hazards-of-the-cloud-data-storage-services-crash-sets-back-researchers/52571?cid=wc
- http://petapixel.com/2014/07/31/cautionary-tale-bug-dropbox-permanently-deleted-8000-photos/
- http://gawker.com/5625139/grad-students-thesis-dreams-on-stolen-laptop
- http://www.plymouthherald.co.uk/Stolen-laptop-dissertation-stored-tomorrow/story-16031933-detail/story.html

There is a saying about storage that goes “lots of copies keeps stuff safe”. The idea behind the principle is that even if your main storage system fails, you still have access to your data.

If you have very important data, you may want to keep many copies, but most scientists should follow the 3-2-1 Rule and keep three copies of their files. This rule states that you should have 3 copies of your data in 2 locations on more than 1 type of storage media.

The offsite copy is particularly critical. Many people keep their data and a backup copy on-site, but this doesn’t factor in scenarios where the building floods or burns down (as can happen in a chemistry building) or a natural disaster occurs. Storing a copy of your data off-site can make the recovery process easier if everything local is lost.

While the 3-2-1 Rule mainly concerns redundancy, it’s also a recommendation for variety in that data should not all be stored on one type of hardware. Computer hard drives fail, cloud storage can be disrupted, and CDs will go bad over time; each storage type has its own strengths and weakness so using several types of storage spreads your risk around. So if the first copy your data is on your computer, look for other options for your backups like external hard drives, cloud storage, local server, CDs/DVDs, tape backup, etc. Finally, always keep a local copy of your data if its main storage is in the cloud. Accidents happen, even with well-run cloud storage, so it’s always best to have a copy of your data in your direct control, just in case.

Here’s an example of following the 3-2-1 Rule using resources a research has locally available:

- a copy on my computer (onsite)
- a copy backed up weekly to the office shared drive (onsite)
- a copy backed up automatically to the cloud

The 3-2-1 Rule is simply an interpretation of the old expression, ‘don’t put all of your eggs in one basket.’ This applies not only to the number of copies of your data but also the technology upon which they are stored. With a little bit of planning, it is very easy to ensure that your data are backed up in way that dramatically reduces the risk of total loss.

Adapted from "Rule of 3" by Kristin Briney (http://dataabinitio.com/?p=320), CC-BY.

Backups

Part of following the 3-2-1 Rule means having backups in place. When looking for good backup options, consider the following:

- Any backup is better than none
- Automatic backup is better than manual
- Your work is only as safe as your backup plan
- Check your backups periodically

You should check your backups for two reasons. First, you need to know that they are working properly. A backup that is not working is not a backup at all. You should test your backups once or twice a year and any time you make changes to your backup system. If your data are particularly complex to back up or particularly valuable, considering testing your backups more frequently.

The second reason to test your backups is to know how to restore from backup. You don't want to be learning how to restore from backup when you're already in a panic over losing the main copy of your data. Knowing how to restore from backup ahead of time will make the data recovery process go much more smoothly.

It's a small thing to periodically test restore from backup, but it will give you piece of mind that your data are being properly backed up and that you will be able to recover everything if something happens to your main copy.

Adapted from "Test Your Backups" by Kristin Briney (http://dataabinitio.com/?p=399), CC-BY, and from the data management guide by UWM Libraries (http://guides.library.uwm.edu/data), CC-BY.

3.4 Data Quality Control

Spreadsheet Best Practice

VIDEO: https://www.youtube.com/watch?v=f11-0Ce1i3I

With so many researchers using spreadsheets, spreadsheet best practices are important for managing data well. These best practices emphasize computability over human readability (this may differ from the way you are currently using spreadsheets). The reason is that a computable spreadsheet can be easily reused in many different analysis programs. With many different software programs available for analysis, having a spreadsheet that is portable and reusable allows researchers to do more with their data.

To make spreadsheets computable, the first thing to do is streamline them. This means getting rid of extra formatting like highlighted cells, merged cells, and excess font variations. This formatting does not encode information in a computable way and, as with merged cells, can sometimes even hinder computability. If you are using formatting to convey important information, find a different way to add this information to the spreadsheet or add it to the spreadsheet's documentation.

You should also collapse all of your data down to one large table wherever possible. This maximizes computability. Collapse smaller tables and place charts on a separate page. Spreadsheet pages should only contain data with one row at the top for column labels. You can include some documentation in the spreadsheet itself, but large amounts of documentation should be moved externally, such as in a data dictionary.

Another good practice is to choose good null values. You use a null value where you have no data to report, such as when you could not record a particular measurement for some reason. Null can be represented with a blank space, "null", "NAN", etc. The important thing is to choose a representation and be consistent. Also note that null is not the same as zero. Use "0" in your spreadsheet to denote when you took a measurement and that measurement was actually zero.

Finally, it's a good idea to keep a backup copy of your data in its raw form. That way, if you mess up your analysis, it's easy to revert back to the original data and start a new analysis. It's also a good idea to keep a copy of your spreadsheet in .csv format if you plan use the data in multiple software programs.

3.5 Documentation

Taking Better Notes

Having sufficient documentation is central to making your data usable and reusable. If you don't write things down, you're likely to forget important details over time and not be able to interpret a dataset. This is most apparent for data that needs to be used a year or more after collection, but can also impact the usability of data you acquired last week. In short, you need to know the context of your research data – such as sample information, protocol used, collection method, etc. – in order to use it properly.

All of this context starts with the information you record while collecting data. And for most researchers, this means taking better notes.

Most scientists learn to take good notes in school, but it's always worth having a refresher on this important skill. Good research notes are following:

- Clear and concise
- Legible
- Well organized
- Easy to follow
- Reproducible by someone "skilled in the art"
- Transparent

Basically, someone should be able pick up your notes and be able to tell what you did without asking you for more information.

The problem a lot of people run into is not recording enough information. If you read laboratory notebook guidelines (which were established to help prove patents), they actually say that you should record any and all information relating to you research in your notebook. That includes research ideas, data, when and where you spoke about your research, references to the literature, etc. The more you record in your notebook, the easier it is to follow your train of thought.

It's also recommended to employing headers, tables, and any other tool that helps you avoid having a solid block of text. These methods can not only help you better organize your information, but make it easier for you to scan through everything later. And don't forget to record the units on any measurements!

Overall, there is no silver bullet to make you notes better. Rather, you should focus on taking thorough notes and practice good note taking skills. It also helps to have another person look over your notes and give you feedback for clarity. Use whatever methods work best for you so long as you are taking complete notes.

Adapted from "Taking Better Notes" by Kristin Briney (http://dataabinitio.com/?p=542), CC-BY.

Data Dictionaries

VIDEO: https://www.youtube.com/watch?v=Fe3i9qyqPjo

Best practices say that spreadsheets should contain only one large data table with short variable names at the top of each column, which doesn't leave room to describe the formatting and meaning of the spreadsheet's contents. This information is important, especially if you are trying to use someone else's data, but it honestly doesn't belong in the spreadsheet.

So how do you give context to a spreadsheet's contents? The answer is a data dictionary.

So what is a data dictionary? A data dictionary is an external document that gives necessary context to a dataset. Generally, a data dictionary includes an overall description of the data along with more detailed descriptions of each variable, such as:

- Variable name
- Variable meaning
- Variable units
- Variable format
- Variable coding values and meanings
- Known issues with the data (systematic errors, missing values, etc.)
- Relationship to other variables
- Null value indicator
- Anything else someone needs to know to better understand the data

This list represents the types of things you would want to know when faced with an unknown dataset. A data dictionary repeats this list (or a variation of this list) for every variable in the dataset to give a full picture of the data.

Not only is a data dictionary incredibly useful if you're sharing a dataset, but it's also useful if you plan to reuse a dataset in the future or you are working with a very large dataset. Basically, if there's a chance you won't remember the details about a spreadsheet or never knew them in the first place, a data dictionary is needed.

Adapted from "Data Dictionaries" by Kristin Briney (http://dataabinitio.com/?p=454), CC-BY.

Templates

Templates are a great way to add structure to research notes and make sure that you've recorded all of the necessary information. This will help you find information later and ensure that no important details are missing from your notes.

So how do templates work? Basically, you sit down at the start of data collection and make a list of all the information that you have to record each time you acquire a particular dataset. Then you use this as a checklist whenever you collect that type of data. That's it.

You can use templates as a worksheet or just keep a print out by your computer or in the front of your research notebook, whatever works best for you. Basically, you just want to have the template around to remind you of what to record about your data.

Let's look at an example. Here's a template that a spectroscopist may use when recording her data:

- Date
- Experiment
- Scan number
- Laser beam powers
- Laser beam wavelengths
- Sample concentration
- Calibration factors, like timing and beam size

Using this list as a template may result in notes like the following:

- 2010-06-05
- UV pump/visible probe transient absorption spectroscopy
- Scan #3
- 5 mW UV, visible beam is too weak to measure accurately
- 266 nm UV, ~400-1000 nm visible
- 5 mMol trans-stilbene in hexane
- UV beam is 4 microns, visible beam is 3 microns

Remembering to record the necessary details is the biggest benefit of using a template, as this is an easy mistake to make in documentation. Templates can also help you sort through handwritten notes if you always put the same information in the same place on a notebook page. Basically, templates are a way to add consistency to often chaotic research notes.

Adapted from "Templates" by Kristin Briney (http://dataabinitio.com/?p=531), CC-BY.

README.txt Files

README.txt files are one of the best data management tools. The reason is that many of us keep notes separate from our digital data files, so our digital data is not always well documented or understandable at a glance. README.txt files cover this gap and allow you to add notes about the organization and content of your digital files and folders. This helps collaborators and your future-self navigate through your data.

README.txt files originated with computer code, where it is the first file someone should look at in order to understand the code (as implied by the name, README). Being a .txt file makes this information readable on a number of systems because of the simple file type. The simplicity and portability make README's a great tool to coopt for data management.

It's strongly recommended to use a README.txt file at the top level of your project folder to explain the purpose of the project, the relevant summary and contact details, and general organization of your files. This is equivalent to using the first page of your laboratory notebook to give a general description of your project.

Here is an example of a top-level README.txt file for an imaginary chemistry project:

> Project: Beth's important chemistry project
> Date: June 2013-April 2014
> Description: Description of my awesome project here
> Funder: Department of Energy, grant no: XXXXXX
> Contact: Beth Smith, beth@myemail.com
>
> ORGANIZATION
>
> - All files live in the 'ImportantProject' folder, with content organized into subfolders as follows:
> - 'RawData': All raw data goes into this folder, with subfolders organized by date
> - 'AnalyzedData': Data analysis files
> - 'PaperDrafts': Draft of paper, including text, figures, outlines, reference library, etc.
> - 'Documentation': Scanned copies of my written research notes and other research notes
> - 'Miscellaneous': Other information that relates to this project
>
> NAMING
>
> Raw data files will be named as follows:
>
> - "YYYYMMDD_experiment_sample_ExpNum" (ex: "20140224_UVVis_KMnO4_2.csv")
>
> STORAGE
>
> All files will be stored on my computer and backed up daily to the shared department server. I will also keep a backup copy in the cloud.

If you hand someone this project folder, the README.txt contains enough information to understand the project and do basic navigation through the subfolders. Plus, the file tells you where all of the copies of the data live if one should accidentally be lost. While not extensive, this information is invaluable to someone unfamiliar with this work trying to find and use the files, such as a boss or coworker.

Besides having one top-level README.txt file, it's also good to use these text files throughout your digital file structure whenever you need them. If you cannot tell, at a glance, what all of the files and subfolder contain, you should create a README.txt (and possibly rename your files and folders!).

Here is an example of a low-level README.txt, which documents the differences between several different versions of analyzed dataset:

Description of files in the “Analysis/ReactionTime/KMnO4” folder

- KMnO4rxn_v01: Organizing raw data into one spreadsheet
- KMnO4rxn_v02: Trying out first-order reaction rate
- KMnO4rxn_v03: Trying out second-order reaction rate
- KMnO4rxn_v04: Revert back to v02/first-order fitting and refining analysis
- KMnO4rxn_FINAL: Final fit and numbers for reaction rate

The graphs corresponding to each file version are in the ‘Graphs’ subfolder, with correspondence explained by the README.txt contained therein.

You can see that README’s don’t have to be large files. Instead, they just need to contain enough information to know what you’re looking at.

README.txt files are ostensibly for other people who might use your data, but they are also useful for you, the data creator, if and when you come back to an older set of data. We tend to forget small details over time and a good README.txt serves as a reminder about those details and an easy way to reacclimate ourselves with our older data.

Adapted from “README.txt” by Kristin Briney (http://dataabinitio.com/?p=378), CC-BY.

3.6 Data Sharing

Data Citation

[VIDEO: https://www.youtube.com/watch?v=He96Mt0o00o]

Whenever you use someone else's data, you should cite them. This is important for helping others understand and follow up further on a research topic, as well as respecting others' efforts accounting for research integrity. Data citation is akin to article citation in that you put the information the "works cited" portion of your report or article. The only big difference between citing an article and citing a dataset is the citation format itself.

While individual citation styles (MLA, Chicago style, etc.) may offer a preferred format, the general format for a dataset citation is as follows:

> Creator (Publication Year): Title. Publisher. Identifier

The identifier is preferably a DOI (a digital object identifier, which is more permanent than a URL) but a URL also works. Here is an example dataset citation:

> Piwowar HA, Vision TJ (2013) Data from: Data reuse and the open data citation advantage. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.781pv

Where available, you can also add information on series, version, and access date to the citation. Any other information should be given in your write up when you explain how you processed the dataset.

Finally, if the data corresponds to a published article, it's good practice to cite both the article and the dataset, especially if you read the article to better understand the data. The most important thing is to be clear about the resources you used by citing them.

Managing Re-use Permissions with Creative Commons

When using someone else's data, in addition to citing them, it is also important to be aware and respect any licensing terms for re-use. Generally speaking, almost all current creative works are considered under copyright law. There are some grey areas concerning research data and any associated documentation that describes the creative act of the original experiment. Additionally, there may be ethical considerations of confidentiality if the data are associated with human and/or animal studies (http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html). When in doubt, consider someone else's material as under rights protection, even if they are freely available on the Internet, and may require permission to re-use and attribution (i.e. citation). Seeking and getting permission can be straightforward for many published materials (http://www.copyright.com/rightsholders/rightslink-permissions/) or involve a complicated process of licensing terms and review.

Creative Commons is a tool for streaming licensing and re-use of digital materials that are openly available on the Internet (http://creativecommons.org/licenses/). It layers a "Common Deed", designation icon and machine readable code over the legal framework that enables creators, human and machine users to quickly determine clear permissions for copying, distributing, commercial or non-commercial uses, attribution requirements and other types of activities. There are 6 variations of the basic license, ranging in permissions from unrestricted dissemination and use, including tweaking or remixing, with attribution; to downloading, sharing and attributing only without changing or commercial use (see chart). Some of the licenses that allow changes to the work require that the user use the same CC licensing terms on any further works, called ShareAlike. There is also a CC0 "No Rights Reserve" designation, which indicates that a work is completely free of rights restrictions and may be shared, changed and re-used without further permission. All materials licensed under Creative Commons are openly and freely available to at least download.

THE CC LICENSES

cc BY CC BY: distribute, remix, tweak, and build upon the work, even commercially, as long as credit given for the original creation	cc BY NC CC BY-NC: remix, tweak, and build upon the work non-commercially, must acknowledge credit for the original creation
cc BY SA CC BY-SA: remix, tweak, and build upon the work even for commercial purposes, as long as credit given and new creations licensed under identical terms	cc BY NC SA CC BY-NC-SA: remix, tweak, and build upon the work non-commercially, as long as credit given and new creations licensed under identical terms
cc BY ND CC BY-ND: redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit	cc BY NC ND CC BY-NC-ND: only allowing others to download and share as long as credit given, can't change in any way or use commercially
cc ZERO CC0: rights-holders waived restrictions so others may freely build upon, enhance and reuse the work for any purposes	0 PUBLIC DOMAIN (Note: not technically a license) Public Domain Mark: works already free of known copyright and database restrictions and in the public domain throughout the world

Adapted from "About The Licenses" by Creative Commons (http://creativecommons.org/licenses/), CC-BY; and "About CC0" by Creative Commons (https://creativecommons.org/about/cc0), CC BY

CHECKING FOR RE-USE PERMISSIONS

What do you want to do with the work and what do you need to make sure you do for this type of use?

	ALLOWANCES	
REQUIREMENTS	Derivatives OK	Commercial Use OK
Attribution Only	cc BY (CC BY) cc BY NC (CC BY-NC)	cc BY (CC BY) cc BY ND (CC BY-ND)
Same Licensing Terms	cc BY SA CC BY-SA) cc BY NC SA (CC BY-NC-SA)	cc BY SA (CC BY-SA)
None	cc ZERO (CC0) 0 PUBLIC DOMAIN (Public Domain)	cc ZERO (CC0) 0 PUBLIC DOMAIN (Public Domain)

Note: all CC Licenses allow for open copy and distribution; CC BY-NC-ND gives no other allowances and does not appear on this chart

The lecture materials in the OLCC are generally covered under a CC XX license [link to OLCC post, TBD]. Several of the materials in this lecture were adapted from others works under the CC BY license. Students may choose to license their course output as well. Creative Commons provides a simple License Chooser tool (http://creativecommons.org/choose/). The most important points to consider are if you want to allow for commercial use, if you want to allow adaptations to your work, and if you want any further works to be under the same license terms. You can provide a URL to the original source work and more information about yourself for easier attribution and include the icon code for your chosen license on the web page for your material. Questions can be directed to: https://wiki.creativecommons.org/wiki/Frequently_Asked_Questions .