

How To Archive Your Thesis/Dissertation/Project Data

Joseph C. Slater*
Associate Professor, Wright State University

June 4, 2009

Contents

1	Introduction	2
2	Data storage	2
2.1	How to store your data	2
2.2	Documenting your data	3
3	Archiving your results: File Formats	3
3.1	Plotting	4
3.2	File formats for graphics	4
3.3	Photographs	4
3.4	Line art, plots, graphs, computer renderings of objects with computer drawn axes/lines, text	5
3.5	Drawings	7
3.6	Combined	7
4	Archiving Data: Storage	7
4.1	Archiving Media	7
5	Usage of Graphics in Documents	8
5.1	Care of Media, Transcription, and Redundancy	9
6	Summary	10
7	Bibliography	10

*Thanks to Frank Ciarallo, Gary Gray and Steven Page for improvements to this document.

1 Introduction

You've spent days/months/years collecting/analyzing your research topic. You put the requested information into your paper/presentation, turn it in, and viola, you're done, right? Not really. What you've just generated is an electronic document of which the only remaining part may be a hard copy in a few years. Further, without all of the splendid data that you used to generate your charts they are now most likely uncorrectable.

This is the case for a large portion of the theses and project reports that students and professionals develop. Many times the only result of a significant amount of effort is a digital photograph of a specific representation of the data. Changing the representation of the data in the future isn't a simple task unless a few simple steps were taken to assure their permanency. Further, setting your graphics in the wrong format significantly degrades the presentation of the effort on paper, on screen, or both. This short document illustrates simple steps that must be taken in order to archive your work for future access.

2 Data storage

Plotting your data and saving the plot is not archiving the data. It is converting it to a visual representation, one that you chose. It may not be the one that you, or the people you send it to, want in the future. Archiving means saving the raw data. Every amount of data processing that moves away from raw data destroys information that is almost always irreversible. Calculated quantities such as mean, variance, integrals, etc, do not retain all of the information in the raw data. Further, saving only those quantities raises an error in their calculation into a monumental blunder should their calculation have somehow been faulty. It is important to archive your data in the least processed form possible. If the processing is very time consuming, archiving the processed result is also a good idea. However, rarely is the cost of archiving raw data excessive.

2.1 How to store your data

Data should always be stored in raw ASCII (American Standard Code for Information Interchange), or raw text, files. Other formats can *additionally* be used. However, those formats may become unreadable in 10–20 years. The ASCII format has been around for a long time, and is the format of the programming languages used to write the codes that read other formats, so it's a good bet that this format will exist indefinitely.

Data should be stored in column delimited format. You can use whatever character you like for delimiting. Spaces, commas, tabs, all can work fine, but make sure that none of those characters are part of your data. Almost all computer codes can save to an ASCII file of some type. This file will almost always be larger than when using other formats. Further, your program may complain that you will lose information in saving in this format. Often, this is useless information. However, try to find out what it is. In many cases you

should also save in native format for the code. In fact, you should almost always save in *both* formats. The ASCII format is your emergency backup for when you can no longer read a data file written by a code that has either a) been extinct for many years, or b) is unavailable to you due to licenses at your new place of employment.¹ If there is a *standard* format for your field, you should also save in that format. This means up to three archived data files in the place of one. If this seems like a lot, then you don't think much of your data, and this document is irrelevant.

2.2 Documenting your data

Just having your data stored in files is not enough. It must be documented. This documentation may be in your report/thesis. If so, this document must be part of the data archive. Further, this documentation must be provided in ASCII format too! We really don't know if MS Word 2000® will be readable in the year 2020. Probably not. If that's the case, when you come back to this data in the future to see if some new processing technique yields new and exciting insight, you'd better hope that the data is documented in an ASCII file. Typically, this file is named "README.TXT". The reason should be readily apparent. This is the first document you should look at. Sets of data described by this document should be in the same directory as this document, or subdirectories beneath it. This document must contain:

1. A description of the experiment, including equipment used (down to serial numbers!), names of files corresponding to pictures useful for recreating the experiment, and enough information to find final published documents using this data (the actual citation is preferred). If the data is the output of a code, a version number and the input information should be supplied. If at all possible, the code itself should be supplied, as well as the compiler requirements (e.g. version 4.0 of MATLAB®, or L^AT_EX, or MS Excel 95®)
2. Nomenclature for directories/subdirectories and files so that the circumstances (case information) of each file can be ascertained.
3. Your name and where you think you are going. Some way to contact you.
4. Names of codes that can read this data (for each file type). This can also be complex. As much as possible, presume the reader has no clue what to do to use this data. Most likely, you will be the reader in the future.

3 Archiving your results: File Formats

The results of your collected data are the computational analysis and resulting plots and figures resulting from them. No doubt this took a great deal of effort. If all of it was

¹Using GNU software protects you from much of this by being free for use in any environment.

generated manually, then you have made life hard on yourself. Most tasks involving data processing plotting are easily automated with modern tools. In performing many tasks by hand, you require the future user of the data to know each and every step you followed to process then plot the data. Guess what, this is the second part of the README.TXT file. Most likely, a user will want to change one or two simple steps of the process. The more complex the process, the more specific the instructions will be.

3.1 Plotting

Languages such as Octave, Gnuplot, MATLAB®, Mathematica®, IDL®, etc allow scripting. A compiled language can also be used with the same results. You should provide a file that when executed will generate the files that became your plots. If the data processing is extensive, the processing and plotting (or rendering) should be separated into two or more tasks, with two or more scripts. Ideally these scripts should not prompt the user for information, but define answers to pertinent questions as variables that can be edited. Better yet is to have one script for each data file, with the names easily identifiable. For example: plotfrdata.m and frfdata.txt .

These files need not read directly from the ASCII file, but should make clear how to load data from the text file if necessary.

3.2 File formats for graphics

Use of the wrong graphics formats for data presentation is the best way to make good results look sloppy. The following is an incomplete list of graphics file formats. It lists the most reliable ones for the tasks stated. Reliable means commonly acceptable and readable.

3.3 Photographs

TIFF (Tagged Image File Format) is a very, very high quality format for graphics such as scanned images and photographs. It also results in some very large files. Joint Photographic Experts Group (JPEG or JPG) allows significant compression with minimal loss of the same types of data, however the compression can lose some information and thus jpeg shouldn't be used when image processing may later be performed. Both work well when various shades of color/brightness are used, with relatively soft edges. Storing a photograph in any other format likely makes no sense at all. Using these formats for graphics other than scanned images or photographs is not recommended. A very coarse description of these formats are that for each pixel they record the appropriate color and darkness to plot. JPG uses schemes to compress this data. In the compression process, JPG (before 2000, which is far from standard) uses techniques which cause strange aberrations in data with lines or text.

1. Preferred archive format; TIFF, JPG (high quality)

2. Preferred presentation format: JPG

3.4 Line art, plots, graphs, computer renderings of objects with computer drawn axes/lines, text

Vector graphic formats are ideal for these cases. Encapsulated PostScript (EPS) is the best way to go, with Portable Document Format (PDF) a close second. Windows Meta File (WMF) may work if you are confident that you are getting vector graphics. Encapsulated PostScript (EPS) or Portable Document Format (PDF) done correctly doesn't know what a pixel is. Thus it prints to infinite resolution by telling the output device simply that there is a line between two locations (not pixels) and that the output device should adjust pixels in its resolution to make it as perfect as possible. This differs from JPG/TIFF which can alias and look quite bad as they are viewed closer or blown up in size. EPS and PDF can be scaled up or down without loss in quality. Generating a JPG or Graphics Interchange Format (GIF) and converting it to EPS is *not* equivalent and not recommended. All of the vector information is destroyed in generating the JPG/TIFF/PNG (Portable Network Graphic)/GIF and cannot be recovered. Screen captures currently also result in poor quality graphics. There are some exceptions. If you think you have the exception, print your graphic at 100 times the capture size. If the text looks jagged to your eye at any level on a 300 dpi printer, you don't have a vector graphic.

For example, observe figures 1 and 2. They are sections of complex graphics. Figure 1 was saved as a JPG, then imported and edited in an unknown package. Figure 2 is the one I regenerated. Note the vastly improved cleanliness of the line. By saving as a vector graphic, the line in Figure 2 is now editable by hand or using an editor such as Adobe Illustrator®. Figure 1 cannot easily be edited to improve the appearance of the line.

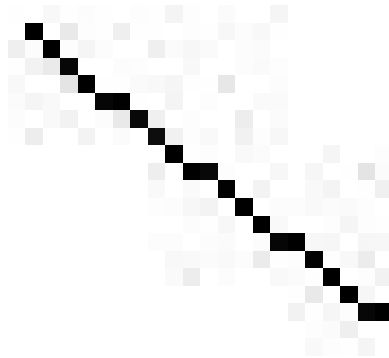


Figure 1: Zoomed line from a JPG file. Zoom in a factor of 10 times and notice the jaggedness of the line.

The website <http://vectormagic.stanford.edu/> can convert pixelized graphics to vector graphics, but there's no substitute for the real thing.

Table 1: Some commands for generating EPS files

Language	Command	Help
MATLAB®	print -depsc2 filename	(help print)
Mathematica®	Export["filename.eps", "eps"]	Help menu
Octave	print	requires octave forge
Generic MS Windows®	Install postscript printer, use this driver to print to a file	You may need to strip out the leading lines. File should start with "%!". Apple Laserwriter recommended.
Generic Macintosh®	Print:PDF:Save as PDF	PDF is standard graphic format on Mac. Converting to EPS requires Adobe Distiller®, Ghostscript, or other converter.

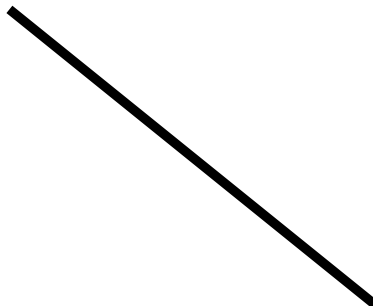


Figure 2: Zoomed line from an EPS file. Zoom in as much as you want. The line will be no more jagged than the resolution of your screen.

3.5 Drawings

The PNG and GIF formats are ideal for graphics with swaths of uniform colors. PNG is preferred by some because there are no license restrictions on the format, but GIF is more common. However, the licenses and concerns that some had for the GIF format ended with the expiration of all patents on the technology in August, 2006. Either of these do well, resulting in high quality graphics with small file sizes. The problem with them is that lines and text can also alias. In most cases it is best to simply use EPS in these cases as well. EPS stores the drawing as a set of objects that can be edited as an object instead of as a painting. PNG and GIF do not understand objects so much as pixels.

3.6 Combined

When you have text on a photograph, what do you do? Use EPS again (or PDF). The JPG or TIFF is embedded in the file as an object, and the text is embedded on top of it. With these formats you can combine the best of both worlds.

4 Archiving Data: Storage

Archiving is more than how the data is organized. It also included physically where the data (the actual storage media) is put and under what conditions it is stored. Choice of media is of critical importance. All of the preceding advice is eventually worthless if the media on which you store your information fails. In my observation most computer users who are prepared for media failure have suffered from failure before. Those who haven't often have too much faith that their hard drive/tapes/floppies/CDs will last forever.

4.1 Archiving Media

Magnetic-based media such as floppy disks and hard drives have a half life of between five and seven years depending on storage conditions. Technically, that means the strength of the magnetic fields drop by 50% in that time. Practically, it means there is a significant chance of data loss in that time. Optical media tends to be more robust, but varies depending on multiple factors.

Floppy disks tend to fair the worst because they are typically manufactured as a commodity item where quality is of minimal concern. This wasn't always the case, but it certainly is now. Floppy disk drives are also manufactured cheaply and their read/write quality varies. Any data currently on floppy disks (zip disks included) should be expected to last no more than a couple of years reliably (if the data is critical, this is insufficient). My advice is to either throw out the floppy disks, or copy the information to a better medium, then throw out the disks. My experience with old floppy disks is mixed.

Hard drives tend to fair better, being mostly reliable for the four or five year life of the typical computer. However, storing something on a hard drive on your computer is not

archiving. You can archive to an external hard drive, but it is critical that the hard drive not be accessible again except in an emergency. It cannot remain accessible electronically to any computer. If it does, it is vulnerable to viruses, power surges, user errors, etc. Hard drives do not provide sufficient long term reliability given that their cost exceeds more reliable methods.

Tapes tend to be the most reliable of magnetic media. Life expectancies of 10-30 years are common, but only under ideal conditions. The National Bureau of Standards publication, Care and Handling of Computer Magnetic Storage Media, recommends that magnetic tape be stored at 65 +/- 3 degrees Fahrenheit and 40% +/- 5% Relative Humidity.[1] For whatever reason, I've observed that tapes are rarely used by individuals for archiving. This was perhaps erroneous until the advent of optical media.

Optical media can be much more reliable, but it is hard to produce, and the cheaper media is not more reliable than magnetic media. CDs/DVDs can fade over time. Life of the media depends on the sealing method, reflective layer, organic dye used, and user storage conditions. Re-writable media should no be used for archiving as they are specifically designed to be changeable after the first burn, i.e. the "burn is not permanent". Thus they are also more vulnerable to decay over time and more susceptible to damage. Further, different formats, e.g. CD-R, CD+R, DVD-R and DVD+R use unique algorithms/concepts for tracking data on the media and providing redundancy. DVD+R is the most robust format, coming after all of the others, and being designed to overcome the shortcomings of the others. McFarland has a rather long and detailed document on this, and recommends using only Taiyo Yuden DVD+R at the current time, available from the SuperMediaStore.com, purported to be the only online US distributor that guarantees that their media is certified Taiyo Yuden.[2] The Taiyo Yuden FAQ also contains useful information.[3] Although this citation is a blog, given the depth of discussion and responses, I believe it warrants respect. Unfortunately I haven't looked for alternative citations on this topic. However, my recollection of articles during the development of DVD writable media agree with the substantial part of this document.

5 Usage of Graphics in Documents

Shortly after reading this document, you will find out that MS PowerPoint® will not accept EPS files for show. They may print. I don't know. However, since that's not the intended use of MS PowerPoint®, it doesn't matter. My answer to that is that this document covers *archiving*, not usage. Although there is overlap, and the information and explanations within still apply, practical limitations always take precedence. For the sake of usage in MS PowerPoint®, I convert my EPS files to PNG or GIF format. In the future, JPG2000 format may be a better choice, but compatibility is a major issue when generating a presentation.

5.1 Care of Media, Transcription, and Redundancy

Media has a finite storage life. Proper care of the media can extend that life. I currently don't have the answers for storage of all forms of media, and am unlikely to ever have them. It behooves the reader to determine what the optimal storage environment is for their media. As most media is designed for use in a typical office environment, it isn't a great leap to presume that room temperature with approximately 40-60% humidity is likely near optimal. This is in fact the case for the media that I've researched (tapes). All storage media should also be kept in the dark (light is energy, can be converted to heat, which causes diffusion) and in as clean an environment as possible. Even for media that can be cleaned, dirt can not cause more good than harm. It can eventually end up in places it shouldn't me.

Further, magnetic media by its very nature should not be exposed to magnetic fields. Fortunately magnetic fields drop rapidly away from the source so keeping the media at least 2" away from a magnetic field in a typical office should be sufficient. It is typically easy to guarantee multiple feet, making any nearby field irrelevant.

Often the technology that writes to and reads from the media is more finite than the storage life of the media. The 5 $\frac{1}{4}$ " and 3" floppy combined had a technology life of approximately 20 years total before being replaced by newer formats with higher storage densities. Many older scientists and engineers have boxes of programs on punch cards that can no longer be read.² As new technology becomes available, it is prudent to occasionally, say every 10 years, create an additional archive of the data in the new format (transcription). It is prudent to understand if that format is useful for archiving or not (long life, inexpensive, sufficiently popular that there will be readers in the future, sufficiently robust to errors). The original archive should not be discarded. The benefit of discarding it (minimal) is likely heavily outweighed by the cost necessary to verify that the new media is a perfect duplicate, and will also survive as long.

Further, as archived data is important, so is its physical protection. Buildings burn down, are flooded, robbed, boxes are lost, ... accidents happen. It is prudent to have a second archive in another physical location. Rarely will this ever be needed. However, if the data is important, it is certainly worth the redundancy. Most people can use their home and office for such locations. I've had to use secondary archives due to failed magnetic media in the past. When one loses data it can be very trying.

While this document is intended to give some insight as to how to archive data, it is not intended to cover backups, which are somewhat like short-term archives that can be replaced. It is left to the reader to investigate the importance of backing up. As a quick summary, weekly backups to an external drive, or better yet, an online backup system (automatically off-site), is prudent. Depending on the rate of your work, quarterly or so archives are also prudent. Just as in archiving, the average user tends to lack sufficient

²Ironically, it is only recently that technology is developing electronic storage media with lives longer than cards!

paranoia.

6 Summary

It's your data, and your graphics. I haven't found anyone yet who has disagreed with this advice, and most have immediately told stories of countless hours wasted compensating for a lack of proper archiving practice. Data must be stored and documented in a meaningful way that is absolutely clear. Regarding graphics, use EPS or PDF for all graphics that are not a pure photographs. Converting a photograph to a JPEG or TIFF is a simple task on a Mac® using the Preview application. On other platforms there is free software available, or Adobe Acrobat® can be used. This document is focused more on “what” than “how.” Modern internet search engines are more than adequate for finding out how to perform these tasks. While this document is far from exhaustive, it will set you on the right path to protecting your hard work for future use.

7 Bibliography

References

- [1] Bogart, J. W. C. V., “A Letter to the Editor of the Scientific American Regarding: ‘Ensuring the Longevity of Digital Documents’,” *Scientific American*, 1995. Available from: <http://palimpsest.stanford.edu/bytopic/electronic-records/electronic-storage-media/bogart.html>.
- [2] McFarland, P., “How To Choose CD/DVD Archival Media,” Available from: <http://adterrasperaspera.com/blog/2006/10/30/how-to-choose-cddvd-archival-media/>.
- [3] CDFreaks, “The Taiyo Yuden FAQ,” Available from: <http://club.cdfreaks.com/showthread.php?t=178622>.
- [4] Kuhn, M. G., “Effective scientific electronic publishing,” <http://www.cl.cam.ac.uk/~mgk25/publ-tips/>.