The Workflow of Data Analysis Using Stata

J. SCOTT LONG
Departments of Sociology and Statistics
Indiana University–Bloomington

A Stata Press Publication
StataCorp LP
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-047-5
ISBN-13: 978-1-59718-047-4

8 Protecting your files

A recent story in the New York Times (Schwartz 2008) began "LISTEN. Do you hear it? The bits are dying." The same day at the National Science Foundation, a review panel evaluated proposals for the $100 million DataNet program, an initiative to develop the next generation of tools for the preservation and access of scientific data. As digital data accumulate, people are discovering to their chagrin that digital data can be more fragile than paper and easier to misplace. The 10-year-old snapshot is still in the album, but the hard drive with last summer's JPGs crashed and the pictures were lost. I know exactly which binder holds analyses from a paper published in 1978, but I need to search several hard drives for the unprinted log files from a paper published two years ago. In 30 years, will it be as easy to access those digital logs as it is to read the printed output from 1978? If I left the printout on a shelf for 30 more years, someone could probably still read them to confirm my findings. Will the same be true for my digital log files? Will it be possible to use a USB drive? Will files be destroyed by a hardware failure or corrupted by a virus? Will newer software read the old file formats? Such concerns affect everything that is stored digitally.

Preservation of files is a critical part of the workflow of data analysis. You want to prevent the loss of files that you are actively working on, to maintain files from completed work that you might need later, and to preserve critical data and analyses for future generations of scientists. My goal is to help you develop a realistic plan for protecting your files. A successful plan needs to consider both the risk of losing a file and the probability of following the plan:

$$\begin{aligned}\Pr(\text{File loss}) = &\{\Pr(\text{File loss using plan}) \times \Pr(\text{Follow the plan})\} \\ &+ \{\Pr(\text{File loss without using plan}) \times \Pr(\text{Ignore the plan})\}\end{aligned}$$

A simpler plan that is religiously followed is more likely to protect your data than a more elaborate plan that is inconsistently applied. Accordingly, the workflow I suggest balances ease of use with the degree of protection. The procedures are simple enough to fit into my daily work and robust enough that my chances of data loss are small. The workflow illustrates general principles and provides a framework that you can adapt to your needs and your willingness to take an active part in backing up files. Fortunately, with current technologies, you can have a very effective workflow for protecting files that requires only minimal effort. The key is to have a comprehensive plan, keep files organized, and automate the process.

The chapter begins by distinguishing among levels of protection that range from simply making a duplicate copy to preserving an important dataset for the next 100

years. Deciding what form of protection you need involves considering the trade-off between the cost of losing files and the cost of preserving files. The most basic tool for preserving files is to have multiple, duplicate copies, but the best thing is to not lose files in the first place. Accordingly, I discuss how files are lost and ways to minimize risks. To provide a context for understanding why any workflow for data protection must include redundancies and anticipate unlikely events, I discuss Murphy's law and what it implies for protecting files. Next I outline a basic workflow designed for a "typical" data analyst who is more interested in substantive analysis than data preservation. Although this workflow might not deal with all your requirements, it provides a basic structure that you can adapt. The chapter ends by discussing long-term, archival preservation.

Before proceeding, I must add three warnings. First, I hesitate to recommend how you should protect your files because these files are valuable and I cannot be responsible for their loss, but file preservation is too important in the workflow of data analysis to ignore. Although the suggestions in this chapter have worked for me, I cannot guarantee that they will prevent you from losing files. Second, before you rely on any method to protect your files, you should test it with files that are not critical. Do not abandon your current method until you are sure that the new method works. Third, the technology for data preservation is changing rapidly. By the time you read this, there could be better ways to protect your files. The Workflow web site will add new information as it becomes available.

8.1 Levels of protection and types of files

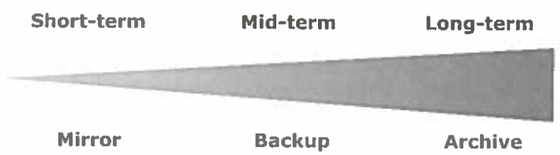

Figure 8.1. Levels of protection for files

The workflow that I present considers three levels of protections (see figure 8.1). Short-term protection focuses on making sure that the files you are using today will be there tomorrow. Protection involves continuously making duplicate copies of your files, referred to a mirroring. Short-term protection protects against the failure of your hard drive, a computer virus, or accidental deletion. Mid-term protection involves protecting files that you have finished working on but that you might want to use in the years to come. I refer to this as backup. As technology changes, you will need to move these files to new media and data formats, but you are not concerned with preserving the files beyond your own interest in them. Long-term protection, referred to as archiving, seeks to maintain the information in perpetuity. Archiving is extremely difficult, requiring

constant concern with migrating files to new media and formats and the availability of documentation that is accessible to anyone interested in the files.

Deciding on a workflow for file preservation requires evaluating the costs of losing files against the costs of preserving files. If a file is mistakenly deleted, how much time and money are needed to replace the file? If you cannot replace it, what are the consequences? Although you never want to lose files, how much time and money can you invest protecting them? Answers to these questions vary for different types of files and depend on the type of work you do. Table 8.1 summarizes costs for four major classes of files.

Table 8.1. Issues related to backup from the perspective of a data analyst

	Type of File			
	Working	Posted	Archival	System
How do you recover a lost file?	Redo recent work	Redo older work	Download files	Reinstall the software
What is the cost of recovering the file?	Minor delays	Potentially lots of work	Minor inconvenience	Minor inconvenience
For how long are you preserving the file?	1–3 years	3–10+ years	100+ years	1–3 years
How difficult is it to preserve the file?	Some work	More work	(Substantial)	Trivial
Should you consider media & format?	Very little	Some	(Critical)	Very little

Note: Items in parenthesis apply if you are concerned with archival preservation.

Working files include documents, do-files, log files, datasets, reprints, and other files that are part of ongoing analyses. Because you are actively working on these files, if they are lost you could probably redo the work. Rather than risking the lost time to redo your work, it makes sense to protect these files. Because the files change every day you work, you need to make duplicate copies of these files frequently. However, because you are concerned only with short-term protection, you do not need to worry about storage media becoming obsolete or file formats becoming unreadable.

Posted files are working files that you have completed and will no longer change. For example, after you create a dataset and verify that it is accurate, the dataset and do-files should be posted. If you later find problems, new files are created but the old ones are left as is. Because posted files are essential for replicating your work, you need to preserve these files for longer than working files. Accordingly, you may eventually need to migrate the posted file to newer storage media and file formats. For critical files, you might want to store the information in more than one format. For example, datasets could be saved in Stata format, ASCII, and SAS XPORT, which is the U.S. Food and Drug Administrations standard. SAS XPORT files can be created with Stata's `fdasave` and read with `fdause`. I would also treat user-written ado-files installed in the `PLUS` and `PERSONAL` folders (page 349) as posted files. These files are not part of official Stata, and you will need them to replicate your work.

Archival files include datasets, codebooks, articles, and other files obtained from organizations committed to preserving and distributing digital information. For example, ICPSR distributes and preserves datasets in the social sciences, whereas JSTOR preserves and distributes published research. If you lose an archival file, it should be easy to obtain another copy. Although you might want to back up these files along with your posted files, this is a matter of convenience rather than necessity.

System files include the operating system (e.g., Mac OS, Windows) and software (e.g., Stata, Stat/Transfer). These files can be replaced by reinstalling the source DVDs or by downloading the files. If one file is corrupted or deleted, it is a minor inconvenience to reinstall the program. If the boot drive of your computer fails, the cost is greater. If you want to recover quickly from losing your boot drive, you can back up the drive including hidden files using utilities such as *Norton Ghost* for Windows or *Carbon Copy Cloner* for Mac OS. Unless you work under strict deadlines, the time required to back up system files might not be worth it. The only system files I back up are for software that is essential for reading critical files or for reproducing analyses.

8.2 Causes of data loss and issues in recovering a file

Although the workflow presented below protects files by making duplicate copies, you want to avoid losing files in the first place, a topic that is now considered.

Deleted and lost files

Files can be lost if they are mistakenly deleted. Most operating systems let you recover deleted files, a feature called Recycle Bin in Windows and Trash in Mac OS X, as long as you make the recovery before the original space used by the file is used by newer files. The `\- Hold then delete` folder suggested in chapter 2 also serves this function but is more dependable because you have complete control over when files disappear. More elaborately, Mac OS X 10.5 includes Time Machine, which backs up files on a daily basis, allowing you to "go back in time" to recover a file that was deleted. This works well as long as your backups fit on one drive and that drive does not malfunction.

A file can also be literally lost—it is not deleted, but you cannot find it. This is the proverbial needle in a haystack, a growing problem as people accumulate more files on larger and cheaper hard drives. The way to avoid misplaced files is to keep things organized, carefully choose filenames, and use a utility that searches by name and content.

You can also lose files if you think they are backed up, but they are not. Before relying on the backups made by network administrators, verify how long the backups are kept and what the procedures are for recovering files. These backups might only be available to restore files after a catastrophic hardware failure soon after the backup was made, rather than to retrieve a single file. Further, some organizations are destroying backups because they might contain sensitive information that could be used for identity theft. If you rely on backups made by others, verify how long the backups are kept and what the procedures are for file recovery.

Corrupted files

A file is corrupted when the information within the file is stored incorrectly due to a write error when recording the data, to deterioration of the media (e.g., a disk goes bad), or a virus. Even one incorrect bit in a 100-megabyte file can make the file unreadable. There are several ways to prevent this problem. First, when a file is copied, do a bit comparison to verify that the source file and copy are exactly the same (see page 337 for further information). When you disconnect a USB or Firewire drive (e.g., an external hard drive, a memory stick), eject it as recommended by your operating system (e.g., right-click on the drive icon and select eject) instead of simply pulling out the plug. Local IT staff tell me that this is the most common reason they encounter for people losing files. Keep your virus software up to date. Files can also become corrupted as media age. Files on a CD left in the sun may last only a few weeks. Hard drives should be replaced after five years. Because files can be corrupted by write errors caused by power fluctuations, use an uninterruptible power supply (UPS) if you live in an area where there are frequent brown-outs or power failures.

Hardware failures

Data are also lost or sometimes corrupted when hardware fails. Failure rates follow a "bathtub curve" with high rates of early-life failures followed by lower rates until the hard drive begins to wear out and failure rates increase. New drives can fail because of a flaw in their manufacturer, as I recently learned when the boot drive on my 3-month-old computer died. Old drives fail because they wear out. Although hard drives are typically rated with a mean time to failure (MTTF) of over one million hours (114 years), Schroeder and Gibson (2007) found observed rates in the range of 2–4% per year.

There are several simple things that help prevent hardware failure. First, turn off your computer by exiting all programs and shutting down the computer as suggested by the operating system rather than simply turning off the power. Second, use a surge protector. If you have unstable electrical power, use an uninterruptible power supply (UPS). Third, make sure your computer has plenty of ventilation and remove dust from the fan; however, a recent study at Google (Pinheiro, Weber, and Barroso 2007) found that heat is not as much of a problem as previously reported. Fourth, do not move your computer when the hard drive is reading or writing. Although hard drives are remarkably robust, if you bump a drive at the wrong time, you can lose a file or the entire disk. Finally, if a drive develops a hum or squeal, replace it immediately.

Obsolete media and formats

Obsolete media and formats are an easy way to lose access to your files. Although this is not a concern with working files, files that have not been used for even a few years can be at risk. My recent experience illustrates the problem. When analyzing data from a 10-year-old study of patients, we discovered that our dataset was missing a critical variable. After a lengthy search, we determined that the file with the missing variable was on a backup tape. Because we no longer had a drive that could read the tape (i.e., obsolete media), we hired a firm specializing in this problem. They recovered the file for $1,000. Then we discovered that the file was in a format no longer supported by current software (i.e., an obsolete format). It took a $1,000 in staff time and months of delay to reinstall an old operating system and the software needed to read the file.

If you do not have access to the equipment needed to read your storage media, the files on those media are effectively lost. Once common media, such as ZIP disks, can disappear quickly. To prevent such loss, you need to migrate your files from older storage media to new ones as technology changes. When you get a new computer, make sure that you can still access the media used with your old computer. If not, find a way to transfer the files before getting rid of the old computer. To gain a better appreciation on how rapidly media appear and become obsolete, look at the online *Chamber of Horrors: Obsolete and Endangered Media* (Kenny and McGovern 2003–2007).

Even if you can copy a file from the backup media to your computer, you need software that can decode the file. It does no good to "preserve the bits" if you have lost their content. For example, if a dataset is in the once common OSIRIS format, but your software cannot read OSIRIS, the data are lost until you find a program that can decode the files. To illustrate the magnitude of the problem, a recent story from BBC News (2007) reported that the British Museum is at risk of losing 508,000 encyclopedias worth of digital information stored in formats that are no longer commercially supported. There is no easy solution to this problem because archival formats have not been fully established. To prevent this type of loss, store critical files in several formats. For datasets, save data in Stata format, but also in ASCII, SAS Transport format using `fdasave` (because it is now the standard for the FDA), and perhaps a few other formats. For text files, save your file in the format you are using (e.g., Word's `.doc`) but also in other formats such as Rich Text Format or PDF.

Recovering lost files

Even if you are careful, you are likely to lose a file sometime. When this happens, hopefully you have a backup copy. But, sometimes a file is deleted or corrupted without a backup. When this happens, take your time so that in your rush to recover the file you do not delete more files or make it harder to recover the lost file. If you have deleted a file that has not been backed up, do not write new files to the drive until you have tried to recover the file. Adding new files can make it impossible to recover a deleted file. If you lose a file because of a problem with your computer or a possible virus, do not connect your backup drive, disk, or tape until you are sure what the problem is.

A colleague once lost a file to a software malfunction and then lost the backup when the same software corrupted the backup. Unless you are sure what the problem is, ask your IT support staff for help. If all else fails, you can use a commercial data-recovery service (search the web for "data recovery"). These businesses specialize in recovering data from damaged disks but are very expensive and most require payment even if they fail to recover your files.

8.3 Murphy's law and rules for copying files

Any workflow to protect files must take Murphy's law very seriously. Murphy's law states: "If anything can go wrong, it will." Much can be learned from the context in which this famous law emerged.[1] In 1949, at Edwards Air Force Base, Murphy's Law was named after Captain Edward A. Murphy who was an engineer working on a project to see how much sudden deceleration a person could stand in a crash. Dr. John Paul Stapp, who rode a sled that created 40Gs of deceleration and lived, noted that the good safety record on the project was due to a fundamental belief in Murphy's Law and extraordinary efforts to circumvent it. He coined Stapp's Ironical Paradox: "The universal aptitude for ineptitude makes any human accomplishment an incredible miracle." When it comes to saving your work, expect things to go wrong, expect that you will delete the wrong file at the worst possible time, and expect a hose to be left on in the room above your computer. If you expect the worst, you might be able to prevent it. Based on this notion and the sometimes painful experience of others, the following rules should be followed no matter what approach you take to preserving your files.

Rule 1: Make at least two copies

Have at least two copies plus the original of all files.

Rule 2: Store the copies at different locations

Many disasters affect everything at a location, so store copies in different buildings. External drives can be stolen, so keeping all backup drives in the same room is bad idea. Similarly, fire and water damage often affect everything in a room.[2]

Rule 3: Verify that copies are exact duplicates

When copying files, use software that does a bit comparison of the source files and the copies. Most copy programs do not verify that the source file and the copy are

1. My history of the law is based on http://www.murphys-laws.com/murphy/murphy-true.html, which reproduces the article "Murphy's Law" from the March 3, 1978 issue of *Desert Wings*.
2. Disasters can occur in places that seem very well protected. ICPSR recently suffered damage when maintenance left a hose running on the floor above their server room (M. Gutmann 2005, pers. comm.)

exactly the same. Without bit verification, you might think you have an exact copy of your file when in fact it is a corrupted version of the file.

8.4 A workflow for file protection

I suggest a two-part workflow for protecting your files, as illustrated in figure 8.2.

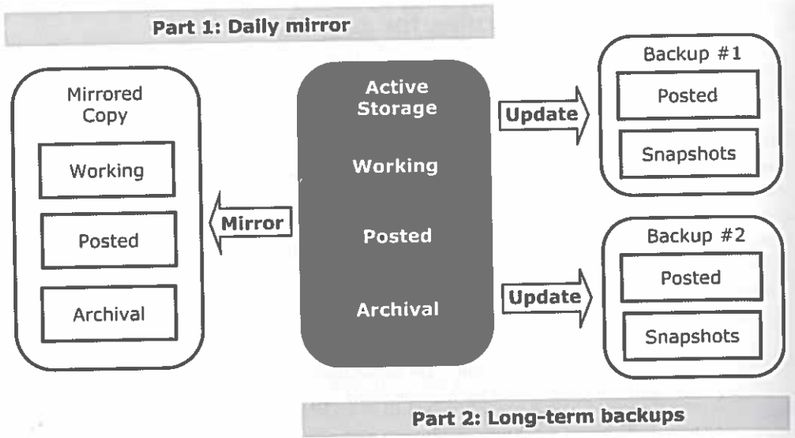

Figure 8.2. A two-part workflow for protecting files used in data analysis

Each rounded rectangle is a storage device. For simplicity, I refer to these as hard drives, but they could be other storage devices. The rectangles within each drive represent the types of files I want to protect. The shaded rectangle represents my active storage. These are the files that I have created and that I am using today. Part 1 of the workflow creates duplicate copies of active storage, referred to as a mirror, to provide short-term protection. Part 2 provides mid-term protection of posted and project files by making two copies of the files, referred to as backups.

Part 1: Mirroring active storage

For active files, the greatest risk is the catastrophic loss of all files due to hardware failure, a virus, or accidental deletion. To protect against such loss, I save duplicate copies of the files at least once every day. The duplicate files are referred to as a mirrored copy because all changes to files in active storage are reflected in the mirror. There are two basic approaches to creating a mirrored copy. With continuous mirroring, changes to files in active storage are transferred to the mirror within a few seconds.

The disadvantage of continuous mirroring is that if I accidentally delete a file, it is immediately deleted on the mirror so that I cannot use the mirror to recover the file. More sophisticated programs for mirroring files move copies of changed files to a separate folder as a fail-safe. Periodic mirroring occurs only at times you select. You might update the mirror as the last thing you do each day before logging off. Because you only change the mirror periodically, you can recover files that changed since the last update to the mirror. The disadvantage is that if the drive with active storage crashes, your mirror will not include copies of files that changed since the last update of the mirror. If you want the best of both continuous and periodic mirroring, you can create both a continuous mirror and a periodic mirror, although this increases the complexity of the workflow.

Configurations for mirroring active storage

The specific way that you mirror active storage depends on how many computers you use, what storage devices are convenient, and the software you use. Here are some alternative configurations, all of which provide excellent protection.

Single computer mirror: If you always work on the same computer, it is easiest to keep your active files on an internal hard drive and to mirror files to an external drive, a LAN, or other storage connected to the computer.

Multiple computer mirror with LAN: If you use two or more computers that are connected to the same LAN, you can keep active files on the LAN and use the internal hard drives on each computer to hold mirrored copies of the active files from the LAN. Because you are using multiple computers, you have extra security since files are mirrored multiple places. If the computers you use do not allow you to save files to the internal drive (e.g., in a public computing lab), you can create a mirror on an external or portable drive that you carry with you. Alternatively, you can use the next configuration.

Multiple computer mirror with portable drive: The best solution for the way I work is to use a portable drive for active storage and carry the portable with me as I move among computers. For years, I transferred new or changed files between computers using a ZIP disk or memory stick but regularly forgot to transfer the file I needed or was confused about which computer had the latest version of a file. By using a portable drive for active files and the internal drives on each computer to hold my mirror, I always have the latest version of my files (unless I forget the portable). On each computer's internal drive, I have a directory `\- Portable mirror` that is used for a continuous mirror when the portable is attached to that computer. When I leave work, I eject the portable drive, put it in my briefcase, and take normal precautions not to drop the briefcase. At home, I plug in the portable drive and follow the same procedures. When traveling, I take the portable drive with me and make a mirror to the internal laptop drive. If I travel without a laptop, I take the portable drive with me and use a memory stick to make duplicates of new files I create.

Multiple computer mirror using the Internet: If you use multiple computers that are on the Internet, you can use Microsoft's Foldershare or a similar program to mirror files over the Internet. Suppose you use one computer at work and another at home. When using the computer at work, Foldershare continuously synchronizes your active files to your home computer. When you are home, you use the files on that computer as your active storage and Foldershare mirrors changes to your work computer. I have experimented with Foldershare, and it is easy to use and efficient with a few exceptions. First, the beta software limits you to 10,000 files per library and 10 libraries. Second, if the Internet connection goes down, the other computer is turned off, or the software has a problem, you need to work without a mirror until the problem is fixed. Third, if you use a laptop that is not always connected to the Internet, you must remember to update files to the laptop before you use it. Fourth, if you work on computers that do not have Foldershare installed, you will not have access to your files.

Other configurations can work equally well depending on how you work and what technology you have available.

Part 2: Offline backups

The second part of the workflow makes backup copies of posted files and snapshots (defined below) of files from active projects.

Posted files

Posted files include the do-files, log files, datasets, text files, and others that you have used in your writing or shared with others so that they should no longer be changed. The rule for posted files is simple: Once a file is posted, it should never be changed. See chapter 2 (page 22) and chapter 5 (page 125) for further details on posting. While I continue to work on the project, I keep posted files from that project in my active storage, but because these files are part of results that I have distributed or shared with coauthors, I want to give them an extra level of protection against hardware failures, viruses, or accidental deletion. About once a month or when I finish a major task, I copy newly posted files from active storage to a backup drive. If my archival files are not very large, I often back them up along with the posted files.

The key to backing up posted files is to automate the process so that the backup software can determine which files need to be copied. If you must manually determine which files need to be copied, your workflow will not work unless you are very patient and exceptionally thorough. The easiest method I have found is to put posted files in subdirectories named `\Posted`. Suppose that I have the following directories on my computer (where for simplicity I use a simple directory structure):


```
\Projects
    \COGA
        \Posted
        \Work
    \EPSL
        \Posted
        \Work
\Workflow
    \Posted
    \Work
```

My backup program only copies files from `\Posted` directories and automatically reproduces the directory structure from active storage:

```
\Posted files
    \Projects
        \COGA
            \Posted
        \EPSL
            \Posted
    \Workflow
        \Posted
```

If I add another `\Posted` folder to my active storage, the next time I run the backup utility it will create the same folder within `\Posted` and will copy the newly posted files.[3]

This plan for backing up posted files only works if your backup program can automatically select all files located in directories named `\Posted`. On my computer, it takes my backup software about a minute to search 4,000 folders to find which files need to be backed up. If your software cannot select files to back up in this way (and not all backup programs can), you need to organize your files some other way that makes it easy to determine which files to back up. For example, you could create one folder, `\Posted`, and create subdirectories for each project. I do not find this approach convenient, but you might.

Before proceeding, I want to emphasize that this plan only works if you stick to the posting principle: Once a file is posted it cannot be changed. For years, I made the mistake of changing, renaming, or relocating "posted files". Then I needed to back up the files again. What about the files I had backed up before with different names? If I deleted them, old do-files would not work and my research logs would point to a file that no longer exists. If I did not delete them, I had multiple files with the same information but different names. This process wasted time and was not reliable. The solution was to follow the simple principle for posting a file: After a file is posted, it can no longer be changed. If I find a mistake in a posted file, I can create a new file with a new name, but I cannot change the original file. If I decide that I no longer need a posted file, I can delete it but cannot change it.

3. The software I use can copy only files in directories that contain a particular phrase, such as "posted". Accordingly, it will copy files from `\Posted`, `\Text-posted`, etc, but not from `\Work`. I also use the rule that if a directory name ends with + (e.g., `\Data+`), that directory is treated as posted. See the Workflow web site for details on software.

Snapshots

I also periodically take "snapshots" of all files, posted or not, associated with a project. For example, when I completed the first draft of the workflow book on January 14, 2007, I backed up all the files in `\Workflow` to `\Snapshots\Workflow\2007-01-14`. If I later need a file that was used for this draft but that was not a posted file and was later deleted (e.g., a figure I did not use), I check here. I also take snapshots before I start cleaning files at the end of a project. That way if I make a mistake, I can easily recover the file. My snapshot folder might look like this:

```
\Snapshots
    \Workflow
        \2006-12-12
            (copy of all workflow files on this day)
        \2007-01-14
            (copy of all workflow files on this day)
    \EPSL
        \2004-06-02
            (copy of all epsl files on this day)
        . . . . . .
```

I might keep the snapshots for a long time or delete the files after a few weeks.

Although I keep at least two copies of files located on different devices stored in different locations, I do not worry about keeping the different drives continuously synchronized. Although keeping them synchronized would be better, I find that this is impractical. Instead, I synchronize the drives every few months. When a backup drive is full, I buy a new drive. With the rapidly increasing capacity of hard drives, the new drive is usually at least twice as large so it can hold all my current files and have capacity for at least that many more files. Using bit verification, I copy all files from the current device to the new one. The old drive is stored as "deep backup".

Configurations for backup storage

You should have at least two copies of each file stored on media located in different physical locations to protect against local disasters such as water damage or theft. The drives are only plugged in when making copies or recovering files, so they are not vulnerable to power surges or viruses. Here are some configurations you can consider for your backup storage.

Backups using external drives. My preferred approach is to use multiple external drives to hold the two backup copies of files. These drives are inexpensive, have fast read/write speeds, and are easy to move. When I want to synchronize the drives, I carry the drive I store at home to work and bring it back home after the two drives are synchronized.

Backups using a LAN. If you have sufficient space on your LAN, you can store one copy on the LAN and the other on an external drive or an internal drive (assuming that drive is not used for active storage).

Backups using enterprise mass storage. If your organization provides enterprise storage, such as tape backups, one copy could be stored there and the other copy on another device.

Backups on the Internet. One backup copy could be saved over the Internet with a company that sells storage. As of mid-2008, "unlimited" storage costs less than $5 a month. The catch is that transfer speeds limit you to copying a maximum of about 7 GB of files a day (i.e., two weeks to copy 100 GB, one year to copy 2.5 TB). After initial copies are made, only changed files are transferred. If you have a hard disk failure, it could take weeks to restore files over the Internet. You can also pay to have DVDs made and mailed. The software for these sites only allows simple criteria for selecting files, so selecting only files in `\Posted` folders would not be possible. To find companies that sell Internet storage, do a web search of "Internet storage", "online backup", or "online storage". Keep in mind that if the company selling storages goes out of business, you could lose your backups.

Backups on DVDs. DVDs have the advantage that once you write to them, the files cannot be deleted (assuming you avoid the multiple-write DVDs, which provide more expensive and less stable storage). DVDs, however, are slow, they hold only about 5 GB so you end up with a lot of disks to organize, and quality DVDs are no longer cheaper per GB than hard drives.

8.5 Archival preservation

Important files get lost, sometimes surprisingly so.[4] On November 22, 1963, President John F. Kennedy was assassinated. A few days later, the National Opinion Research Center (NORC) conducted a national opinion survey known as the Kennedy Assassination Study (KAS). After the tragedy of September 11, 2001, NORC decided that the KAS survey should be replicated to allow comparisons of these two tragedies. The KAS codebook was found in the NORC library, a new survey was constructed, and interviews began on September 13. The original KAS data could not, however, be found in the NORC archives, nor were they found in other data archives, such as the Roper Center and the Inter-university Consortium for Political and Social Research. The search then began for the original 80-column punch cards, with hopes that they would be found in NORC's 24,000 cubic feet of storage. Three thousand, six hundred and sixty-nine boxes of cards were listed on the storage inventory, but KAS was not on the list. Someone noticed that while the inventory listed the contents of 3,669 boxes of cards, there were 8,348 boxes of cards in storage. A retired staff member remembered a memo from 1987 that included bar codes for the KAS cards. Amazingly, the 11 boxes of cards were found. Next NORC had to locate a card reader. A firm in New York was hired and the cards were hand delivered. The card reader was literally a little rusty, causing further delays. When the cards were finally read, the multiple-punch data on the cards were

4. This example is based on Smith and Forstrom (2001).

not correctly interpreted.[5] Two months later, the data were decoded. Unfortunately, the variable names simply indicated the card number and column position (e.g., `c3c14`) and no value or variable labels were available. Adding this information caused further delays. Smith and Forstrom (2001, 14) summed up the experience:

> The lessons from the KAS experience are simple but important. Survey data must be sent to survey archives like the Roper Center and ICPSR where the documentation and data will be preserved, backed-up, periodically updated as technologies change, indexed, and made routinely and easily accessible to researchers. Failure to archive studies is poor science and a disservice to other contemporary researchers and those in the future.

A second example of lost data also involves a defining event in modern history.[6] On July 20, 1969, half a billion people watched Neil Armstrong walk on the moon. Macey (2006) summarized the event: "The heart-stopping moments when Neil Armstrong took his first tentative steps onto another world are defining images of the 20th century: grainy, fuzzy, unforgettable." Indeed, the images were of poor quality, obtained by pointing a television camera at the monitor that was receiving the transmissions from the moon. The original images were of too high a resolution to show on TV. Over the next 30 years, the moon tapes moved around, although where they ended up is uncertain. In 2002, a technician from Australia's Honeysuckle Creek ground station found a tape in his garage that seemed to be from the moon landing. Although his tape was not of the walk on the moon, it initiated a search for the moon tapes that had once been stored at the National Records Center. Documents were found showing that 26,000 boxes of tapes had been requested by Goddard during the 1970s and 1980s, but no trace of the tapes was found at Goddard (Kaufman 2007). In 2006, NASA admitted that the tapes were lost. Dolly Perkins who led the failed search explained (Kaufman 2007): "Maybe somebody didn't have the wisdom to realize that the original tapes might be valuable sometime in the future. Certainly, we can look back now and wonder why we didn't have better foresight about this." Following NASA's admission, news stories appeared around the world discussing how NASA lost this historic and scientifically valuable data. Macey (2006) at the *Sydney Morning Herald* wrote a story titled "One giant blunder for mankind: how NASA lost moon pictures." Remarkably, the Australian film producer Peter Clifton heard about the tapes on TV and remembered that in 1979 he had purchased two moon tapes to use in a rock film about Pink Floyd's album *The Dark Side of the Moon* (Egan 2006). His two tapes were recovered from a Sydney vault. In October of 2006, nearly 100 moon tapes were found in the basement of a lecture hall at Curtin University of Technology in Perth, Western Australia (Amalfi 2006). It remains to be seen if these are the right tapes and whether the remaining tape drive at Goddard still works.

5. Multiple punches are a way in which one column of a card can record information on multiple variables. This allows one card to hold much more information but requires special processing to decode the information.
6. This example is based on articles by Macey (2006), Egan (2006), Amalfi (2006), and Kaufman (2007).

I find that most data analysts have not thought much about archival data preservation. I had not until I spent four years on the Council of the Inter-university Consortium of Political and Social Research (ICPSR), a nonprofit organization dedicated to preserving data. When I joined the council, I thought of backing up and archiving as the same thing. I made two copies of files, one on tape and one on disk, and thought I had archived my files. I learned that archival storage is much more complicated, as the two examples above illustrate. The hard part in archiving files is anticipating changes in file formats and physical media, and making sure that the files are documented in a way that is clear to someone who is not associated with the original work. When it comes to archiving, there is no difference between losing a file because it was deleted or losing the knowledge about what the file contains. A good source of information on these issues is *The Guide to Social Science Data Preparation and Archiving* (ICPSR 2005, available at http://www.icpsr.umich.edu/access/dataprep.pdf) and the *Digital Preservation Management Tutorial* (http://icpsr.umich.edu/dpm/).

The best way to archive datasets is to let someone else do it! I highly recommend depositing your original data with an organization that specializes in data preservation, such as ICPSR for the social sciences. For analysis files from published papers, you should consider depositing the files at the journal's archive if they have one (Freese 2007). Not only does this ensure that the data are preserved, but it gives other researchers access to the data for replication and additional research.

You should start thinking about archiving data when you start a project, not at the end of the project. I have projects that I began before I understood what was required to archive data. I kept careful documentation that I could use, but this information was not in a form that others could use. As a consequence, I have data I would like to archive at ICPSR, but it would take a great deal of time to get the data and documentation in a form that others could use. If I had been thinking about preservation from the start, archiving the data would have been much simpler.

8.6 Conclusions

Procedures for preserving the files used in cleaning, analyzing, and presenting your data are a critical part of the workflow of data analysis. Like filing taxes, having your teeth cleaned, changing oil in your car, and removing leaves from your gutters, preserving your files is something you need to do. Most people realize this, but very few people appear to do it systematically. The best way to avoid data loss is to have a comprehensive plan for protecting your files and following it conscientiously. Hopefully, this chapter will help you find a manageable workflow for protecting files and will encourage you to make this a priority. If you adapt the suggestions in this chapter to your own needs, you should be able to do a very good job protecting your files with a minimum of effort. If you begin planning how to archive data as it is being collected, you can save a lot of time and increase the chances that your data will be preserved.

Finally, this chapter focused on procedures for copying files to multiple locations. It assumes that you know what your files are! This seems easy, but versions of files can quickly get out of hand, especially in collaborative projects. In a recent article on data preservation (Schwartz 2008), Dr. Margaret Hedstrom echoed the issues of organization and provenance that I discussed in earlier chapters: "Which architectural drawings of the many versions generated for a project were actually used to erect the building, and what was the chain of decisions that led to the brick-and-mortar result?" Preserving your files is much easier and more effective if you keep files organized and carefully named.