7 “Mostly” Easy Steps to Data Security

by James W Brunt

Data Security for researchers in environmental sciences is an important issue. I’ve tracked many of the issues associated with data security and have blogged about them on occasion. The material below has been compiled from several of these blog posts along with some material accumulated for upcoming blog posts. You may catch me in a contradiction or a conundrum - this is expected. I know of at least two in this document (If you can find them I’ll give you extra credit). This document primarily addresses data security issues faced by the researcher in near-term timeframe but can just as easily be applied to project information managers working at long timeframes. This document **does not address all sensitive data** security issues. Though many of the same steps can be applied to securing sensitive data there are many other guidelines for these data including institutional policies, and state and federal laws. I would probably need at least 13 steps and several days to cover sensitive data.

One of the most frequent questions I am asked by student and veteran researchers alike is “How do I protect my digital data and documents?” When I dig a little I find that most people are doing some kind of irregular backup of their computer files to flash memory drives or USB hard drives. They have an intuitive sense about the value of strong passwords, file naming conventions, and making backup copies. What they are looking for from me is the assurance that they are doing an adequate job of protecting their valuable research products. With this reassurance they can breathe easier and concentrate on the real job of doing science. In this document I will outline some practices for protecting your data. I will not discuss long-term archival or tell you how to backup and restore your computer’s operating system after a disk crash. Instead, I will focus on the daily protection of your data and research works in progress. [Sidebar: I confess to liking odd numbers - thus “7” easy steps instead of 6 or even 4 which would have been far easier. There are ancient and powerful forces at work around odd numbers, especially the number 3 - for example, you know that accidents happen in threes - (they also happen in 1’s and 2’s but we only count the ones that add up to 3).] Near-ubiquitous use of computers and digital storage devices and near-constant Internet connectivity raise many concerns for the safety and security of your digital data. The keys to avoiding unintended consequences to your computer or loss of information are understanding risks and being cautious, aware, and prepared.

1. Know the risks - The highest probability risks to digital data fall into 3 general categories:

- Theft or physical damage to computer or removable media - For example, laptop theft is growing at a rate of 20%/year, over 750,000 laptop thefts annually. According to the FBI, 97% of stolen laptops are never recovered. Any personal electronic device creates an invitation for theft. Smaller devices are at a higher risk of damage.
- Routine Hardware or Software Failure - For example, the average failure rate of a normal computer hard disk 2 yrs old or older, is somewhere between 7 and 10%. That is, 1 out of every 10-14 disks will fail within the year (flash drive failure rates are even

higher). Disk drives are cheap, but the cost of replacing one can be enormous if sudden failure forever separates you from your data and documents.

- Data Tampering or Unauthorized access to data - For example, a Post-Doc was recently caught tampering with a colleague's experimental results. I have not previously included this category of risk when discussing non-sensitive data but recent issues of intra-lab data sabotage and data tampering have moved this risk into the realm of the possible.
- Failure to follow good practice - While loss is not always immediate and unrecoverable nonetheless good practices in file naming and handling and metadata documentation can save enormous energy.

My purpose is not to make you paranoid - just cautious, aware, and prepared. Hopefully the steps below will provide you with non-invasive solutions to give you peace of mind without causing distrust and resentment among colleagues. It's really just a matter of establishing good practice because the consequences of realizing any of these risks may be far reaching and include loss of data, loss of time and effort, damage to professional reputation and position.

2. **Avoid a Notorious Single Point of Failure** - I read almost weekly some new heart rending story about a years worth of data being lost when a dog chewed a thumb drive or a computer was stolen, or a budding author in New York losing his first novel when his computer is run over by a bus! The first rule of 3 on data security is to always have your data in three separate physical places. My solution I call Brunt's Axiom of Here, Near, and There - not sure it will make it into the history books but let's apply this to backing up your research data:

- **Here** - Here is wherever you are generating your content- be it a laptop or desktop computer, portable hard disk, or the Web. This is the source for your digital information, where your work gets done. For most researchers *Here* is a single computer; for some small but growing number *Here* is web-based applications like Google Docs. Other individuals carry their research lives around with them on a USB hard drive and use a variety of available computers. Wherever *Here* is for you, it's the place you store and manipulate your data and documents and it should be reliable and secure.
- **Near** - If you have access to an institutional file server and procedures to store data and documents - use them. You are doubling or perhaps tripling your protection depending on local practice. If not, this is where a nice fat USB hard drive comes in handy. These range in capacity up to terabytes and prices are dropping all the time. In the *Near* is where you have to make some decisions. You can use the operating system backup tools to write regular backups to your USB or Network drives - MAC OSX has a particularly nice one. WINSCP is a particularly powerful tool for MS Windows users that can synchronize between drives. Or, you can do manual, drag and drop, backups and store your files in such a way to make them easy to recover. In general, do not make a flash drive your *Near* data store. Grade A and B flash drives can be safely used for transport and for extra copies if you are not abusing them too much, e.g., carrying them around in a backpack that sometimes sits in a hot car. I want my *Near* data copies close and easily accessible 'when' I make mistakes.
- **There** - In the not to distant past, having copies of data *There* meant sending boxes of disks or tapes to your mother. Software companies now offer online storage solutions

that are cheap, fast, reliable, and intuitive. Some examples are Sugarsync, Mozy, DropBox, Syncplicity, Fasmule, and Opendrive, to name a few. My personal favorite for ease of use and multi-platform support is DropBox. These services basically all work the same - they provide you with a certain amount of disk space out *There* somewhere in the 'cloud' (In the case of Dropbox this storage lives on the Amazon S3) and give you access to it through the web and through a variety of desktop applications. This is where these services begin to sort themselves out for you by offering slightly different features and by supporting different platforms. Dropbox supports web access as well as access to online storage for MAC, PC, Linux, and Smartphone platforms, and it utilizes the folder structure on your platform, so there's no application to open. Dropbox automatically copies any files put into the dropbox folder to the online storage. I keep my work folder inside the Dropbox folder. If I create a Dropbox folder on another computer my files are synced there too.

These approaches can be adapted for alternative computing and research lifestyles but remember the rule of 3.

Activity: set up a free account on one of the available cloud services and use it to meet the requirement of Brunt's Axiom. Try to defeat the safeguards the service offers. For example can you overwrite a file in such a way that you lose a previous version? Be prepared to discuss the benefits and drawbacks of such an approach.

3. Use consistent file naming and versioning - Now you've solved your single-point-of-failure data security problem, having three or more copies of your data creates a data management problem. You'll need directory and file naming conventions. This is a set of personal rules you define for naming data files and the folders you keep them in and for saving multiple versions of files. In general, I like to:

- Keep names short if possible to avoid running into application specific limits.
- Use names that that are at least somewhat descriptive of the contents of the file.
- Include a date using a consistent format (for example YYYY-MM-DD - ISO 8601) that you can later search for.
- Don't include the folder name in the file name.
- If you are going going to be responsible for your own versioning, include a version number at the end of the file name such as v01. Change this version number at consistent milestones (for e.g., each time the file is saved). For the final version, substitute the word FINAL for the version number.

4. Practice safe file handling - To practice safe file handling you'll need to address issues of concurrency, synchronization, versioning, and provenance.

- Concurrency - Concurrency occurs when a file is open for editing at the same time. The potential for concurrency issues is exacerbated by the use of synchronization tools like DropBox.

- Synchronization - Synchronization is necessary when the edited version of a file needs to be copied back so that other copies are the latest version of the file. In the past, this was my biggest headache and resulted in some elaborate programming solutions. Today, technology has caught up with demand, and keeping files in sync is getting easier all the time. If I copy a file to my USB drive to replace one that already exists, the old version is overwritten. If I copy a file to my Dropbox, the program replaces the file while saving the old version. In this way, as I make and document changes to my data, I'm always sourcing the same file knowing that the changes are kept.
- Versioning - Copying or creating a file with an implied record of changes from an existing file is referred to as versioning and most of the online services mentioned above take care of versioning files for you. Versioning problems occur when you have created different parallel versions of a file that need to be merged later. This usually results in a later version overwriting changes in a parallel earlier version. Above we implied that your "Here" copy of your data was the source. If you have many files or are doing a lot of editing you may want to designate a "There" as the source and engage a "checkout and commit" mentality about your data. This is what software engineers do to mitigate these issues. There are technical solutions using software like Sharepoint or source code control systems like SVN, Git, or CVS. These require some commitment to be successful. I do a combination of filename versioning and using software versioning in google docs and DropBox, others I know use version control systems in combination with off-site storage. You'll have to find a system you're comfortable with.
- Track provenance **-** With all this software doing things for you it's hard to imagine that you have anything left to do but there is a manual component that only you can do and that is to document in words the differences in each of the versions. Seeing the changes made in a data file or record is easily done by software but knowing why those changes were made is a job for scientists. This metadata applied to the differences in data is known as provenance. You can apply this information at the file level or at the record level but what you ultimately seek is an unbroken chain of provenance from the source to the most recent version and right on out to the active or archived version. Provenance metadata can be recorded in a very structured fashion or in a narrative form. There are no standards and no good guidelines for recording provenance in the environmental sciences. There are disciplinary approaches that have been used for decades but best practices have yet to emerge. Provenance tracking is an active research area but most research is focused on automated processing systems. So record what you do, when you do it, and why you did it, when processing your data.

Activity: Do a search of the internet using search terms related to tracking provenance in research to get a feel for the depth and breadth of this topic.

5. Practice safe computing - One of the most effective ways you can protect yourself and your data is to take simple, preventive steps with your computer:

- Make sure your operating systems and applications have been updated to the latest version of security updates.

- Make sure the firewall on your laptop is enabled and set to the highest security setting.
- Make sure that you have anti-virus software installed and that it is using up-to-date definitions.
- Make sure you have a full backup before you leave and that it is safely stored away.

With the theft statistics above in mind there are some additional steps you can take to prevent your personal and professional information from falling into the wrong hands when going about with your computer:

- Never lose sight of your computer, particularly at airport screening stations. Put onto the conveyor immediately before passing through screening so that it's more likely to still be in the x-ray machine when you are done.
- Do not put your computer in checked luggage. Use a carry-on bag. If possible use a backpack and not a bag that advertises "computer inside".
- Make sure your computer prompts you for a password when coming out of screensaver, stand-by, hibernate, and after reboot. If you use applications such as the browser to store passwords make sure they are protected by a master password.
- Make sure your computer is labelled and that you have recorded the serial number.

There are now recovery services that you can subscribe to that use software and geolocation to guarantee recovery of your computer should it be stolen. These services range in price from $40-$100 per year and usually include software that will lock the computer and cost recovery insurance as well. An example of this service "LoJack for Laptops" by Absolute Software.

When on the road you will encounter many other dangers not the least of which are the public WiFi networks. A public network is a network to which anyone has access such as those available at airports, hotels, restaurants, coffee shops, even on some airplanes. When you connect to a public network, your online activities can be monitored by others. In addition, malicious individuals may use a plethora of attacks and exploits to gain access to your computer . Some may even operate fake Wi-Fi networks that are designed to fool you into using them in order to attack your system unnoticed.

When possible, use a Wi-Fi network hosted by the hotel or business rather than picking a public Wi-Fi network at random. Some coffee shops and hotels will provide you with an encrypted connection and a key or passphrase that you should take advantage of when available, but this is rare. Even with Wi-Fi encryption, your communications could still be intercepted by other users of the same Wifi network. Always use an encrypted connections from browser or email client when connecting from a public network. Online mail and social network sites such as Gmail, Facebook, and Twitter enforce encrypted connections. This is indicated by the prefix "https" in the browser URL window or a nondescript padlock icon located somewhere on the browser window. For email clients you will need to establish the necessary connection parameters from the hosting organization to ensure an encrypted connection for both sending and receiving of email. These connections are common and fairly standard, but few clients use them by default and may be passing your personal information in readable text every time you connect.

If you are concerned that you cannot securely access the internet via a public network, another option is to connect to the internet via your smartphone 3G or 4G network connection. This is called “tethering.” The approach varies with the phone and may require an additional monthly charge. This approach may be enough to meet your immediate needs and in some cases may far surpass the available bandwidth on public networks. If so, the security provided by the smartphone broadband provider is better than public WiFi.

Don’t use public computers as an alternative to public networks. These are the greatest danger you might face in your travels. Public computers are fraught with perils, the worst of which is the keylogger - a hardware or software based keystroke recorder. This easily installed device records every keystroke a user makes and can be analyzed to produce login names, passwords, bank account numbers, PINs and more. The only preventative for this exploit is avoidance. Limit your public computer activities to non-personal browsing.

6. Use strong passwords - Cutting through the hype and folklore associated with passwords can be daunting and intimidating. You don’t want a brute of an IT lord telling you what to do and not do. Trying to manage the numerous passwords and rule sets required to get through a single week is bad enough.

Despite what you’ve been told, changing passwords frequently is not a panacea - it does protect you but more realistically and importantly it protects entire systems from one particular type of brute force attack. But statistically you are more likely to have your password compromised through a malware attack from a malicious website.

Long, complex passwords are usually more secure than short, simple ones, but they also are more difficult for the user to remember, leading to the increased possibility that users will write them down somewhere, making them visible to prying eyes.

The possible choices for each character of a numerical password are 10 (0 through 9). Possible choices for passwords using letters are 26 for each character. By combining letters, numerals, special characters and upper and lower case, there can be up to 95 possibilities for each character. A four-digit numeric PIN has keyspace of 10,000; that is, there are 10,000 possible combinations. An eight-character password using 95 possibilities for each character has a keyspace of 7 quadrillion. The ability to remember even several dozen of these 7 quadrillion passwords is beyond that of mere mortals, so we fall back on tricks and technology. Mnemonics can be used effectively and password management tools can simplify the job, but they must be properly secured and can create a single point of failure if compromised. Here’s a place where applying a bit of risk management can be useful to reduce the number of personal passwords from dozens to a manageable handful, each of the appropriate strength.

First categorize your passwords based on risk. I work with 3 (surprise) that seems to be the safe path between too few and too many and corresponds nicely to standard risk categories of high, medium, and low.

- High Risk - Passwords that can expose personal information of yourself and/or others. These would be websites that give direct or indirect access to finances or financial data like online banking and bill paying sites, university and institutional portals, commercial sites where you have stored credit card information (Amazon, Itunes), and any site that has your exact date of birth, SSN, or other identifiers that would aid in identity theft.

- Medium Risk- Passwords that expose personal information that is already generally available. These would be websites that may have your contact information for mailing or shipping but no financial information.

- Low Risk - Passwords that expose minimal personal information. These would be websites that require a login to view content like online newspapers, and blogs.

High risk passwords should be strong and unique for each type of information exposed. Every banking or financial website should have a unique password. Consider using a strong password generator and password manager for high risk sites. Medium risk passwords should strong but need not be unique for every site. Don't reuse passwords between categories or sites. Low risk passwords can be anything that does not give away any personal information or clues to higher risk passwords.

Actviity: Test some of your favorite passwords using a secure password tool - for e.g., https://www.microsoft.com/security/pc-security/password-checker.aspx

You can use a password manager to safely store and protect your strong passwords that you can't remember. I prefer KeePass (http://keepass.com) which is again open-source and cross-platform. It uses encrypted storage and also has a strong password generator. It can be stored on and run from a flash drive and can be secured with a keyfile in addition to a password. There are versions for smartphones as well.

Activity: Download and install KeePass on your computer. Read the documentation. See if you can set up a portable KeePass system using a flash drive or other storage. What is the purpose of the "keyfile"? Be prepared to discuss. Use the password generator to generate a strong password. Test this password with the secure password tool above.

7. Be paranoid - in the words of Joseph Heller, in *Catch 22,* "just because you're paranoid doesn't mean they are not after you". Being paranoid about computing security is a bit like 'Pascal's Wager' - even if you don't believe you will ever be a victim of an attack , taking the steps towards securing your personal information is the safe bet. There are a number of things you can do that might seem like overkill and draw looks from colleagues but that will hedge your bet. Some examples include:

Think like a network - recognize the ‘interconnectedness of all things’ when network computing is involved. An innocent thing like you email address can form the critical piece of information the bad guys need to connect information from one source with information from another source where a known social or technical weakness can be exploited.

Use two-factor authentication - for applications in the high-risk category if two-factor authentication is offered you should take advantage of it. Many financial institutions offer this as do large cloud providers like Google. Two-factor authentication involves a second, usually single-use, authentication parameter - this is usually a 5-6 digit code that is generated and sent to your phone. It can take the form of fingerprint scans or facial recognition, smart cards, etc.

Use pseudonyms for login names - using unique pseudonyms for login names to network based services can greatly reduces your network connectivity and thus the risk of being tracked, targeted, and exploited through connections pieced together from social networks.

Don’t accept the ‘default’ - not accepting the default privacy settings on network based applications does two things. First, you fix possibly harmful default information sharing settings and second, you become aware of the level of privacy that you are being asked to agree to. These settings are usually the ones that are preferential to provider - it’s often in there best interest to keep privacy settings loose.

Use data encryption - Data encryption can be a useful technology in securing research data as well as your passwords on your computer or digital storage device. You can encrypt an entire drive, a drive partition, or individual files. You can use built-in operating system file and file-system encryption or/and you can use a number of software tools. When looking for encryption software always look to open-source. It’s important that the implementation of the encryption algorithms be open and reviewable *en masse*. GnuPG and TrueCrypt are my personal favorites because of the broad cross-platform support. I’m not going to attempt to explain the public key infrastructure (PKI) in this document but both of these programs allow you to encrypt without explicitly setting up a PKI.

Activity: Download and install TrueCrypt on your computer. Create an encrypted virtual volume (M:) and a hidden encrypted virtual volume (N:). Practice mounting and unmounting the volumes and putting files in and out. Don’t put critical files in it until you are confident in your understanding of the software.

Summary

Data security is a broad and complex area of interest and importance even when applied to a very narrow window of research data. There is much to learn and need for communities of practice to fill gaps. The best advice I can give is to follow the 3 directives from the introduction, be cautious, be aware, and be prepared. And, in the immortal words of Douglas Adams: “Don’t Panic!”

For the LTER Information Management Short Course - 14 August 2012

The author is a Research Associate Professor at the University of New Mexico and has 25 years of experience in environmental information management and information technology both in the public and private research sectors. He currently serves as the Chief Information Officer for the U.S. Long Term Ecological Research Network. He has written and collaborated on numerous books and articles on the subject.

Use this space for notes and questions