**Module 4: Data Storage, Backup, and Security**

University of Massachusetts at Amherst
MJ Canavan

Steve McGinty

Rebecca Reznik-Zellen

Learner Objectives:

1. Understand why data storage, backup, and security of research data are important

2. Understand data storage, backup, and security methods for research data

3. Understand best practices for research data storage, access control, migration to newer storage media, and security of research data

4. Identify an approach to creating a data storage, backup, and security plan for a project

**Introduction**

Funding agencies look to an applicant’s data storage, backup, and security strategies as an indicator of merit. In dealing with data there are many variables that need to be accounted for. The most meticulously devised methods for the early stages of research can come to naught without a sound plan to store, backup, and secure the data. This module describes those variables and how an applicant can demonstrate readiness for risk.

Storage, backup, and security are interrelated. During the early planning stages of a project, researchers must ensure coordination for these three elements. For example, the choice of hardware for storage must be compatible with the subsequent choices for backup. Both primary and backup storage must have adequate security mechanisms in place. It is a continuum, from hard drives to automatic backup to encryption; the project planning and data management must account for all of it.

**Storage**

Data storage is fundamental to any research project. Without safe, reliable, or accessible storage, your research will not have a home. Storage refers to the media to which you save your data files and software—for example a hard drive, a DVD, or flash drive. Planning for data storage is important because all storage media are vulnerable to risk and will likely become obsolete over time.

Storage media are either optical or magnetic. Optical media include CDs and DVDs; magnetic media include hard drives and tapes. Both are vulnerable to environmental conditions and deterioration over time and should be handled carefully. Data storage for even short-term projects should involve both media types, ensuring access to the data should media fail. Storage media should be updated regularly, between 2 and 5 years.

Using these media types, there are several different options for storing data, each with benefits and cautions to consider.

**Personal Computers and Laptops**

Personal computers and laptops make use of an internal hard drive to store all your data and system files. The internal hard drive is the most immediate storage option that you have. When your computer is functioning normally, you can quickly and reliably access the files on your hard drive. These are convenient for storing your data locally as you are working on it, but should not be the only storage system that you use. Local drives can fail or PCs and laptops may be lost or stolen leading to an inevitable loss of your data.

**Network Storage**

Network storage drives are standalone storage banks that are typically managed by IT staff centrally or within your School or College and accessed using LAN or internet connections. Centrally managed storage is subject to backup and security protocols that may be difficult or time consuming to implement individually. At the same time, centrally managed storage may have restrictions on access or file size that may impact your research. If you are working on a cross-institutional group generating terabytes of data, central storage may not be a good solution for you. It is worth investigating the central storage infrastructure and policies at your institution when considering your data storage needs.

**External Storage Devices**

An external hard drive sits outside of your computer and is connected via data cable. They store large amounts of data, and most allow you to schedule automatic backups of data. External hard drives have many benefits including storage of old files, backing up important data, convenience, storage for copying and transferring. It also provides security, both through encryption and the simple detachment of the drive.

Removable storage devices such as external hard drives, USB flash drives, CDs and DVDs, can seem an attractive option for storing your data due to their low cost, portability, and ease of use. A CD-R (recordable) and CD-RW (rewritable) are types of optical drives that can create CD-ROMs and audio CDs. ROM stands for Read Only Memory; computers can read CD ROMs, but that cannot add content. A feature of many CD-R drives, called multisession recording, enables you to keep adding data to a CD-ROM over time. This is extremely important if you want to use the CD-R drive to create backup CD-ROMs. There is another type of optical drive that can create DVDs. The biggest difference between a CD and DVD is space – a DVD can hold more than 4 times as much data as a CD. However, the longevity of removable media is not guaranteed, especially if they are not stored correctly. In addition, they may not hold enough data, so multiple disks may be necessary (see the UK Data Archive’s [Caring for CDs and DVDs](http://www.bl.uk/blpac/pdf/cd.pdf).) Optical storage is not recommended for long-term storage.

USB flash drives are typically removable and rewritable. USB flash drives offer potential advantages over other portable storage devices. They have a more compact shape, operate faster, hold much more data, can have a durable design, and operate more reliably due to their lack of moving parts. As with any removable storage media, physical labeling is important to identify what data the device(s) contain.

**Remote Storage Services (i.e. The Cloud)**

Remote storage services provide users with an online system for storing and backing-up computer files. Mozy, A-Drive, Microsoft, and Amazon, to name a few, enable users to access data from anywhere with an Internet connection. Using banks of servers located around the globe, these services store and synchronize data files and offer redundant backup services for users. Remote storage solutions are extremely convenient and reasonably priced, but should not be the only storage solution in a data strategy. As third-party providers, a vendor’s terms of service should be examined for indemnification, if the service is terminated, or other constraints on ethical data storage. For example, some data may be required to be stored within U.S. borders, depending upon institutional policies or guidelines for handling sensitive data. See the [US Department of Commerce](http://www.bis.doc.gov/policiesandregulations/advisoryopinions/jan11_2011.pdf) for more information on this issue. Review a cost comparison chart for major cloud services: <http://cloud-services-review.toptenreviews.com/>.

It’s important to be aware of the differences between data sharing services and data storage services. While commercial cloud services like Dropbox™ and Google Drive™ are used often for sharing data, their basic (free) packages have limitations. For example, the basic no-fee Google Drive™ provides users 15 GB of storage. If you are near capacity, Google Drive™ automatically initiates “pruning”, to “trim down” document revisions (See <https://support.google.com/drive/answer/95902>). Once individual revisions have been pruned, they cannot be restored. (There is a way to prevent this, by going to the file menu and clicking “Make a Copy” each time you want to save a revision). When using Dropbox™, it’s important to note that it keeps older versions of files for 30 days and then automatically deletes them (unless you upgrade your account to get the Packrat™ feature). Not all institutions and projects authorize the use of commercial cloud sharing and storage services; check with your institution to see if you are allowed to utilize these types of services.

Commercial cloud storage and sharing services may have size, cost, or privacy limitations that could pose a risk to your data. Be sure to read the fine print and not rely on commercial options solely for storing your data. Commercial web applications can be discontinued unexpectedly, and you will want to know what happens to your data in that scenario. You want to know details about privacy and about how much storage you have, for how long, and for how much money.

**Physical Storage**

Just as you would name any digital data files according to a standard naming convention, labeling is critical for physical storage of data. Any analog materials, from paper hardcopies of survey data to refrigerated lab specimens, should be appropriately labeled with the minimum metadata required to correctly identify the item. This could include creator, date created, associated project and project files, and ownership information. See Module 3 for more information about descriptive metadata.

If moving, shipping, or storing analog materials, appropriate identification of any containers as well as the items within the containers will help to avert confusion if materials are misplaced or mishandled. Maintain a manifest of contents and label data down to the smallest discrete item. Use standard conventions for labeling items that are consistent between digital and analog data.

**Backup**

Backup is an essential component of data management, mitigating the risks of accidental or malicious data loss. Backup allows you to *restore* your data in the event that it is lost. Backup is important for all data, but particularly for research data that is unique or difficult to reproduce.

Examples of data loss:

* Disasters (floods, fires)
* Theft
* Hardware or software malfunctions
* Unauthorized access

A backup strategy is a plan for ensuring the accessibility of research data during the life of a project. What your strategy will be depends on the amount of data you are working with, the frequency that your data changes, and the system requirements for storing and rendering it.

Original

External Local

External Remote

Consider your data storage and backup strategy before you start collecting and creating your data. Your strategy should be able to accommodate the amount of data that you anticipate collecting and be stable for the length of time that you anticipate keeping your data.

**Store at least two copies of your original data**

Storage is the foundation of your backup strategy. Best practice recommends that you store at least two copies of your original or master data files, an external (external to your primary workstation) locally-held copy and an external remote copy, using a combination of the storage media described above. Redundant storage kept in different geographic locations ensures that if a disaster occurs in one place, a copy of your data still exists in another place.

**Create an appropriate backup routine for your project**

After determining where you will store your data, consider what kind of backup procedure you should use. One way of determining this is to consider what would be required to *restore* research data in the event of data loss. Would you need just the data files themselves, the software that created them, or customized scripts written for data analysis? Depending on your research project, you may want to perform full, differential, or cumulative backups.

* **Full:** A full backup will replicate all the files on your computer. Full backups take a long time and require the most storage space. However, full backups are also the most complete and can restore data quickly.
* **Differential Incremental:** A differential incremental backup copies only those files that have changed since the last incremental or full backup. To run differential incremental backups you must first create a full backup as a point of reference. Incremental backups are fast and require the least storage space. However, restoring data using incremental backups is time consuming and requires each differential incremental backup made since the last full backup.
* **Cumulative Incremental:** A cumulative incremental backup copies only those files that have changed since the last full backup. A complete backup is created if no previous backup was done. Using a cumulative incremental backup procedure, you would need only two data sets to restore your files, your last full backup and your last differential backup.

Another important part of a backup strategy is the frequency with which you run backups. If you are making frequent or important changes to your data, you should backup your files on a daily basis. If you modify your data files less frequently, a longer backup schedule – weekly or monthly – may be sufficient. Using the native utilities on your computer (Backup and Restore for Windows, Time Machine for Mac) or third-party or open source applications, you can establish a regular backup schedule for your system and indicate which media the files should be saved to.

*If you are working from a networked computer, your central IT division may already have backup protocol in place. Contact your IT department for their backup plan.*

It is important to estimate the length of time that your data needs to be accessed and preserved and the amount of data that you will need to store over that time. These variables will determine your best choices for storage media and a backup strategy

**Create digital surrogates of analog materials**

If you are working with analog materials, consider making digital surrogates that can serve as backup copies to your original documents. Scanning paper lab notebooks, survey results, notes, or other printed material will ensure that you can *restore* the data in the event of data loss.

**Test your system**

Always test your backup system. Once you have a storage and backup routine in place, go through the exercise of accessing the backup files to be sure that your procedure works and that you will be able to *restore* your data if you need to.

**Other Considerations**

**File Formatting**

Equally important as the media to which you store your data, are the formats in which your data are stored (see <http://www.data-archive.ac.uk/create-manage/format/formats-table>.) Researchers should choose software that is non-proprietary, in an open documented standard, and in common use by the research community. It should also be formatted in a standard representation such as ASCII, Unicode, and PDF, etc. Researchers need to be thinking about these issues as they develop their projects. This should be built in to the project from the outset, not as an afterthought at the end of the data analysis. Consider the questions.

* Who might be using it?
* How will it be used differently in the future?
* Is there a risk of data corruption, missing data or data loss?
* Will there be application performance issues?
* Will there be technical compatibility issues?
* Does migration require downtime?

**Data Migration**

The rapid changes in software and hardware raise compatibility issues. These can arise in a matter of just a few years. The usefulness of the data may diminish considerably if future researchers cannot get access to it. Data migration is the process of translating data from one format to another, either to utilize a new computing system or as a mechanism to preserve data for the very long term.

**Assigning Responsibility for Data Storage and BackUp**

Responsibility for backup and storage of data will often be guided by hardware and software decisions. If a central location or service is selected by the researcher(s), then the individuals in charge of that service should have frequent open communication with all parties. Regularly scheduled backups outlining when, who, and how the backups will occur should be conducted. This is another instance where the data can quite easily be lost to future scholars if these consistent practices are not followed. There are online remote backup services available, but researchers should check with their campus IT.

If researchers and their hardware are more dispersed, then there may be cases where each individual is responsible for his or her own backup and storage. Multiple handling responsibilities can lead to unclear backup and storage plans. Large, cross-institutional or cross-departmental projects with multiple partners creating and managing data would benefit from shared storage and backup strategies with defined roles for all partners.

**Security**

There are different levels of security to consider for your research data.

* Access: This refers to the mechanisms for limiting the availability of your data
* Systems: This covers protecting your hardware and software systems
* Data Integrity: This refers to the mechanisms for ensuring that your data is not manipulated in an unauthorized way

**Protect access to your data**

**Unique User ID/Password**

Unique user IDs ensure that activity can be traced to specific individuals. It is a way to authenticate and authorize access to a server and the data therein. A resource manager program uses unique user IDs for auditing and for checking authorization. User IDs and passwords are assigned to one person and one person only. Password strength is a metric based on how resistant a given password is to being compromised by simple guessing, also known as a brute force attack, or by more sophisticated attacks. Length, complexity, variation, and uniqueness can all contribute to the strength of a password. Adding just one additional character to a password or passphrase makes it an order of magnitude harder to attack via brute force. Use of a phrase that is meaningful only to you is often a good way to generate a unique password that is easily remembered. Adding punctuation such as commas and semicolons within the phrase can further strengthen the pass phrase. Avoid passwords that use dictionary words, sequences or repeated characters, (12345, abcdef, 55555, etc.) and most importantly, personal information such as your name, license numbers, birthday, etc. In addition to the brute force attack, hackers may employ sophisticated tools to decipher passwords. Passwords may be created by a password generator. A password generator is a software program or hardware device that randomly and automatically generates a strong password. Many computer systems include an application that generates random passwords. How passwords are stored in a given system is of critical importance with regard to security. Passwords should be stored in a system that employs some form of encryption. In the event that the server itself is compromised, there is a better chance that the passwords contained within it will not be compromised. The use of a password manager or tool such as Msecure (<https://msevensoftware.com/>), Lastpass ( <https://lastpass.com>), or KeePass (http://keepass.info/) may be advised to help researchers manage their own passwords in a more secure way.

**Access through a centralized system**

Modern data centers and centralized storage solutions are often more reliable than dispersed individual machines. Hard drives are the most burdened and hardworking parts of any computer or server, and therefore no hard drive will work forever. Centralized storage helps users overcome the consequences of damaged hard disks by saving data on several hard disk drives, (HDDs) in redundant arrays (RAID). They can provide backup services, physical and virtual security, environmental controls, restoration services, and scalability. Centralized data centers have dedicated staff that troubleshoot hardware and connectivity issues. If your institution provides a centralized data computer and storage option, determine if its services can meet your data management needs.

**Role-based access rights**

Role specific access levels to data can be assigned. A user account only grants limited privileges to data. Use this type of account for your daily work. An administrator account grants full privileges to make changes. If an unauthorized person gains access to your computer or if the computer is infected by a virus or suffers some other form of attack at the administrator level, then your computer is vulnerable to more damage. Just as a user account restricts what changes you can make to your computer, it also limits what an attacker can do. Limiting administrator privileges, as a safety measure may limit the damage from malicious web sites that try to install software without your knowledge.

**Limitations of wireless devices to protect access**

Mobile devices cannot be protected by the same means as hardwired devices. At present, tablets and smart phones lack the protection provided by antivirus programs and firewalls. The greatest hazard to tablets and smart phone security is loss or theft of the device itself. It is important to set a password for your device. Make sure a PIN is required to access your home screen. Run the latest version of your OS and be sure all patches are applied. Limit the amount of sensitive data you store on your device. Turn off Wi-Fi and Bluetooth if you are not using them. When you do use Wi-Fi, try to be sure you are on an encrypted network requiring a password. Download apps from reputable or official app stores.

**Protect your computer systems**

**Updated anti-virus software**

Virus protection software should be updated daily (this can usually be automated) and should be running in the background continuously. Antivirus software protects against malicious programs including viruses, worms, Trojans and other forms of malware.

**Up-to-date versions of client software and media storage devices**

All key software packages should be updated regularly including OS, browsers, productivity packages, etc. Up-to-date software works in concert with virus protection software and will limit your computers’ vulnerability. All applications that interact directly with the internet are especially vulnerable. Oftentimes these updates will be automated for you as part of an overall network plan if your computer or devices are supported by a centralized IT structure. Patching refers to a fix that is made available by the software creator or trusted 3rd party to correct vulnerabilities or exploits within that software. Some exploits (security holes in software or applications which can be used to gain unauthorized access to a system or network) are called zero day exploits. A zero day exploit is an attack that takes advantage of a security hole before the vulnerability is known. Patches should be applied immediately as time to infection through an exploit can happen very quickly.

**Use of a firewall**

The function of a firewall is to block unwanted network traffic from reaching your computer or server. By limiting network traffic that is not “allowed” you reduce the threat of malevolent intrusion. By limiting unwanted network traffic you limit the potential infections or malware. Firewalls can be both software-based and hardware-based. Windows, Linux, and Mac OS include software-based firewalls. The simplest form of hardware–based firewall is your typical home router.

**Use of intrusion detection software to monitor access**

While a firewall functions to block unwanted traffic while permitting legitimate communications, intrusion detection software detects and alerts administrators about intrusion attempts. It does not block these attempts. Whereas a firewall can be thought of as a security guard that stops unwanted traffic, intrusion detection software is more of a surveillance camera that show you an intrusion in progress.

**Physical access**

Security starts at the physical level. If possible confine your servers to climate controlled, locked rooms accessible by trusted staff and systems administrators. All consoles and dashboards should require authentication to be unlocked. Lock desktops and laptops. Locking lugs prevent theft of hard drives from desktops. Cable locks can secure laptops to furniture. Make sure that office doors are locked. You should be sure to lock your desktop when not in use. Set a time out for your desktop so that the screen locks and will not allow anyone access unless the password is re-entered. This is easy to implement on a variety of platforms. Theft or loss of mobile and portable devices can represent a major security threat.

## Ensure your data’s integrity

### Encryption

Encryption is coding information that cannot be read or deciphered unless someone has the decoding key. Encryption can be used for data in transit (being moved across the network) or data at rest in a storage medium. Excellent discussions of encryption including public and private key cryptography can be found at

<http://searchsecurity.techtarget.com/tip/A-primer-on-encryption>

<http://en.wikipedia.org/wiki/Public-key_cryptography>.

**Electronic signature**

An electronic signature is the electronic equivalent of paper based signing authority. The strongest electronic signature is called the standard electronic signature or digital signature. It is a kind of coded message that is unique to both the document and the signer. A digital signature is meant to ensure the authenticity of the signer and by extension, the document. Changes made to a document after it has been signed in this manner invalidate the signature thereby protecting the document from information tampering. Digital signatures can ensure the authority of the data. A digital signature is the most secure form of electronic signature. It is based on Public Key Infrastructure (PKI) and cannot be altered or tampered with. Public Key Infrastructure (PKI) allows users of unsecure networks, i.e. the Internet, to exchange data securely through the use of a public and private cryptographic key pair. Since 2000, digital signatures carry legal significance. A digital signature (standard electronic signature) seals documents by providing evidence of the user’s identity and the data integrity, and authenticity of the document.

**Watermarking**

Watermarking embeds a digital marker for authorship verification and can alert someone of alterations made to data files. It is most often used with media and images. Watermarking software exists for tampering detections and to embed metadata.

**Activity**

Review the Teaching Case: Regeneration of Functional Heart Tissue in Rats. The following passages illustrate problems of storage and backup. Explain the issues and discuss potential solutions.

* *“At any time, some tissue that was sectioned and on slides may be in one freezer, and some tissue that had not been sectioned yet, but is embedded and ready to be sectioned is in another, and still there is other tissue that may be sitting in a container someplace in another freezer. It should be entered into the spreadsheet saying what was done and where it is, but that doesn’t always happen.”*
* *“The older lab notebooks are in the PI’s office are not backed up. The ones that are currently in use are in the lab. The lab notebook has to be in pen on specific paper because this paper is supposed to be good for 100 years.”*
* *“We back up the data sets on an external hard drive some place. The optical and electronic images are both backed up. The current backup system we are using truncates the data set name to 6 digits then puts a tilde sign and number starting from 001.”*
* *“The files are not password protected or anything. The lab notebook is either in the PI’s office or in the lab, which has key card access.”*

**Data Storage, Backup, and Security Checklist**

* Have you consulted with IT at your institution to determine what resources it provides with regard to storage, backup, and security for your data and research?
* Where will you store your data?
	+ PC or laptop
	+ Removable media (USB)
	+ External hard drives
	+ Network drives
	+ Remote storage (the Cloud)
* Where will you store your backup copies?
	+ PC or laptop
	+ Removable media
	+ External hard drives
	+ Network drives
	+ Remote storage (the Cloud)
* How will you create and sync your backup copies?
	+ Automatic system tools
	+ Manual
* What kind of backup will you run?
	+ Full
	+ Incremental
	+ Differential
* How frequently will you run your backup?
	+ Daily
	+ Weekly
	+ Monthly
* Who will be responsible for running your backup?
	+ Data creators
	+ PI
	+ IT manager
	+ Other
* How will you organize your backups?
	+ Retain file naming conventions
	+ Label external media consistently
* What formats will your files be stored in?
	+ Non-proprietary
	+ Open documented
	+ In common use by the research community
	+ In standard representation (ASCII, Unicode, PDF)
* Have you tested your backup files?
* Does the physical location where your computers, servers, and data storage reside have appropriate security controls?
* Will external media related to your research such as paper lab notebooks be kept secure in locked cabinets with access logs and a list of authorized users?
* How will you protect access to your data?
	+ User ID/password
	+ Limited network access
	+ Role-based access rights
* How will you protect your hard and software systems?
	+ Anti-virus software
	+ A systematic plan for updating and patching all applications and OS
	+ Firewall
	+ Anti-intrusion software
	+ Restricted physical access
* How will you protect the integrity of your data?
	+ Data transferred over the network will be encrypted.
	+ Access to data related to my research is accessible only by those who are authorized to access it.
	+ I have a plan for validating the integrity of my data.

**References**

Storing Your Data, UK Data Archive

<http://www.data-archive.ac.uk/create-manage/storage/store-data>

Backing Up Data from the UK Data Archive

<http://data-archive.ac.uk/create-manage/storage/back-up>

Baylor University Example of a campus IT Backup Page

<http://www.baylor.edu/its/index.php?id=40552>

U. of Edinburgh Information Services. <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/storage-backup>

How to start and keep a laboratory notebook, Jennifer Thomson, *Professor, Department of Molecular and Cell Biology, University of Cape Town, South Africa*

<http://www.iphandbook.org/handbook/ch08/p02/>

RCR Data Acquisition and Management

<http://ori.dhhs.gov/education/products/columbia_wbt/rcr_data/foundation/index.html#3_B>

Wisconsin Alumni Research Foundation, Lab Book Guidelines

<http://www.warf.org/inventors/index.jsp?cid=18>

<http://data-archive.ac.uk/create-manage/storage>

<http://datalib.edina.ac.uk/mantra/storageandsecurity.html>

<http://en.wikipedia.org/wiki/Backup>

<http://www.tech-faq.com/full-backup.html>

<http://www.tech-faq.com/incremental-backup.html>

<http://www.tech-faq.com/differential-backup.html>

<http://www.pcworld.com/article/170688/7_backup_strategies.html>

<http://workawesome.com/general/data-backup-and-sync-1/>

<http://www.guidingtech.com/574/windows-backup-restore-guide/>

<http://www.theverge.com/2013/1/15/3876790/eyebeam-hurricane-sandy-digital-archive-rescue>

<http://www.computerhowtoguide.com/2012/03/why-external-hard-drive-is-important.html>

<http://www.library.umass.edu/services/services-for-faculty/preserving-your-digital-materials/>

<http://www.webopedia.com/TERM/D/data_migration.html>