5-1-2015

Managing the life-cycle of data

DeeAnn Allison
University of Nebraska-Lincoln, dallison1@unl.edu

Managing the life-cycle of data

Data management includes four areas: discovery, preservation, access, and security. Although they don't all need to be mentioned specifically in the data management section of a grant, they do need to be address somewhere in the grant.

Figure 1

Responsibilities for data management

Discovery ensures that the data is adequately described to it can be identified, understood and located. Preservation includes the technical aspects of long-term content management which may require format conversion, data migration to another system, removing data as retention periods expire, managing versions of data, and back-ups and disaster recovery processes. Security ensures that data is safe from unauthorized changes or access, data corruption, and securing data through back-ups and disaster recovery processes. Access is the process of controlling who can view and/or download data. This is customized to the requirements of the data researcher and can include open access, embargoed for specific time-frames, restrictions to individuals or groups, or total restriction.

The type of research will determine how each area is managed, and it is important to ask questions about each area to identify the proper repository. A local repository may be the answer for very restricted data, while an open repository will work well for data that is to be available without restriction. Publications may have their own requirements for data, when researchers plan to submit an article for publication. It is never too early for a researcher to consider these issues.

Preservation

There are several aspects to data preservation: file integrity, data migration, file compression, back-ups, and disaster recovery.

File Compression

File compression reduces the number of bytes in a file to improve file transmission over the Web. It is useful for text files. Image files are already compressed as either lossy or lossless so they do not benefit much from further software compression. Databases are another file type that doesn't compress well, it is better to export the contents into another format, which can then be compressed as needed. Compression can be accomplished either through software or hardware. Software compression takes an original file and reduces the file size by some algorithm that requires an application to restore the file, for example tar and zip files. An archive is another form of compression and consists of a single file that is created out of multiple files, which can also be compressed to save space. Generally it is better not to archive files if space is sufficient because archived applications require specific software for unpacking, which users may not have. It also adds another layer of complexity for long-term preservation of files which must be managed both for the original file type and the archived format. Finally, the content of compressed files is hidden and will not be searchable so the metadata describing these files is critical for discovery.

Even with these disadvantages, large data sets may need to be compressed to save space. In the case of large volumes of scientific and statistical data, searching to find specific instances within the data set may be impractical without exporting and importing the data into another application. In these cases it is acceptable to compress the data sets, but end users will need to be able to transfer the files and use their own software for analysis.

When data is stored in a dark archive (an archive that is not accessible by the public) or on a back-up it can be cost efficient to compress data as a cost savings. When this occurs, procedures need to be in place to periodically test the archive to verify that it is remaining viable.

One of the best techniques for compression isn't a compression method at all, it is de-duplication. Removing duplicate information can save a considerable amount of space. This is especially true when there are multiple versions of information. If it is necessary to keep multiple versions, those reasons should be documented in the metadata.

Hardware compression is built into the repository hardware and operates directly on the disks. It can include de-duplication, virus checking and other management features. All repositories that use hardware compression should be using a lossless compression so that no data is lost. Hardware compression will save disk space for already compressed file like multi-media files.

Unlike software compression, hardware compression is controlled by the repository so depositors don't have any choice except to select a different repository. Whatever compression is selected there is

always an overhead for compressing and decompressing files for delivery to end-users so access speed may be reduced.

File Integrity

File integrity is a device process that validates the integrity of files stored in a repository using a machine coded verification method. This method can include a comparison between the current file state and a baseline state. This method involves comparing the checksum of the file's original baseline with checksum of the current file. Fixity checking verifies that a file has not been altered or corrupted either during transferring or storage. Repositories should run periodic fixity checks to ensure that repository objects haven't been damaged by "bit rot". When corrupted files are identified they can be replaced through backups and other redundant copies.

Data Normalization

Normalized data are data that conforms to common characteristics that can be easily re-used by researchers and read through a variety of readily available applications. We discussed image file formats earlier, but databases create a different set of challenges. Databases are generally normalized to minimize duplication and support consistency. This creates a challenge when the data is exported into a different format like CSV. Metadata that describes how to re-construct the database structure is imperative for the data to be useful outside of the original database. Storing a complete database in an institutional repository is difficult because repositories are also databases. This means the database must become an object in the repository database. Because of this problem, only databases, like Microsoft Access that are discrete files, can be deposited as objects in a data repository. Databases that are stored as objects must be periodically refreshed to remain viable with upgrades to the applications that can load the database. When possible, it is better to export the data into a file format that can be read by other applications, with sufficient accompanying information being provided on interpreting the data correctly.

Back-up and Disaster Recovery

Periodic file back-ups and off-site copies of data form the foundation of data preservation. LOCKSS (http://lockss.standford.edu) or "lots of copies keeps stuff safe" has been a long term strategy for preserving files by simply keeping multiple copies. However, duplication of files without a preservation strategy does not contribute to long-term preservation, and can become a nightmare to manage. When multiple copies are going to be retain, they should be part of an overall plan for data protection that includes periodic checks for file integrity and supports file migration for out-of-date formats. The Digital Preservation Network (http://www.dpn.org) DPN, is another network recently formed to provide a dark archive for preservation.

Disaster mitigation is also an important consideration for data management. All data management plans should include information on data back-up and off-site data duplication to mitigate disasters. Keeping a tape off-site, participating in a preservation network such as LOCKSS or DPN, or using a cloud-based vendor service for storing data can address these issues. Cloud storage is a growing industry in

the US and provides some low cost solutions for storing data. When evaluating any service consider the following issues:

1. Security – who will have access to the data?
2. Storage location – is it stored in the US or off shore. Confidential data should not be stored in a foreign country where control may be lost.
3. What is the method of replication, recovery, and data transferability for when you want to change cloud services.
4. How easy is it to expand capacity, and how much will it cost?
5. What services/support is available from the vendor?

Security

All data repositories must confirm to best practices for security. A discussion of server and network security is beyond the topic for this course, however data security is a topic of interest. Depositors will want assurances that their data is secure and will comply with their requirements for access and retention. Data curators will need to engage researchers with the following questions:

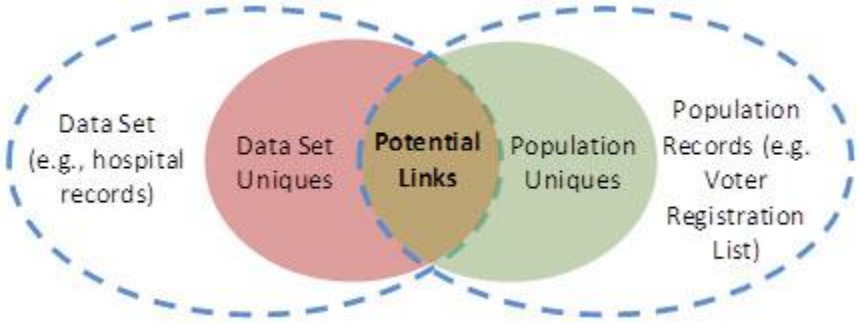
1. Who should be able to access the information? Should there be restrictions based on a login? IP address? Other?
2. Should open public access to the data be provided? Is there an embargo period where no access will be allowed?
3. How long should data should be retained? Permanent retention is not always necessary as sometimes the data will be re-acquired because of improvements in equipment.
4. Will there be multiple versions? If so, what should be retained and for how long?
5. Question researchers who want data retained permanently without public access. Forever is a very long time and it is impossible to guarantee that **no one will ever** view their data. It also seems unreasonable to devote resources to maintain data forever that no one will ever view.
6. Are there copyright, trademark, patent or other restrictions on the data?

In addition, librarians need to be concerned with data that might include identifying information that could violate the privacy of research subjects. All data must be de-identified, or anonymized, so all information that could reveal a research subject's identity must be scrubbed from the data before it is deposited in a repository. Identifying information can be removed from datasets entirely, or coded, or encrypted. Information can also be masked by changing data values or by aggregation. To ensure that the research is complying with local IRB policies, a repository may want to add a copy of the IRB approval with the data.

Data curators should scan data before it is approved for deposit. This includes: checking for social security numbers, gender, names, addresses, phone numbers and other personal identifying information, which can be used alone or in combination to re-identify an individual. Re-identification is the process of taking pieces of data mined from the internet and putting them together to connect information to locate individuals. The US Department of Health and Human services website

(http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html) provides specific recommendations for safe-guarding health information. The following chart from their sites describes the complexity of the situation. Although researchers are already aware of specific information in their dataset, they may not realize that their data can be combined with other public information to reveal individuals identity.

Figure 4



Data curators should evaluate the data with an eye to this type of re-identification. If enough information is retained in data, which a competent person could use to re-identify persons, the librarian should contact the researcher for a discussion about re-identification of data. Although it is the ultimate responsibility of the researcher to protect the privacy of subjects, the librarian is in a position to help advice information researchers on the latest technical developments regarding re-identification of subjects that researchers may not be understand. Unfortunately, this is a moving wall. As technology becomes more sophisticated new possibilities might make it possible to re-identify persons using data that was previously considered safe.

Encryption is used to encode data so that only authorized users with specific credentials can view the data. It can be used to hide sensitive data, but there is an obvious overhead that includes long-term management of encryption methods or keys. It probably makes the most sense in dark archives where data is being placed off-site for storage and where access is limited to the institutional depositor. For publicly accessible archives it should never be relied upon as the sole security measure, and there must be compelling reasons for using encryption that should be documented in an institution's policies.

Access

There are many different types of repositories that archive and provide access to data. The factors that a researcher should consider include:

1. Funder Restrictions or Requirements on Data Sharing
2. Data type (File Format and Quantitative v. Qualitative)
3. Amount of data
4. Potential re-use, importance, or re-analysis of the data
5. Privacy and security needs for the data

There are two types of repositories: institutional and domain. Institutional repositories are supported by the institution, while domain repositories are specific to a field of study, or specialty organization. Institutional repositories are charged with collecting and managing the output of their institutions. Institutional repositories are sponsored by the institution where the researcher works. This can pose complications when the researcher no longer works at the institution. Institutional repositories must have procedures for transferring, receiving transferred data, or indefinite hosting of data after the researcher leaves. When they are largely supported by the institution, they can provide a low-cost alternative to managing data that would not be accepted in other repositories. This includes restrictions on access and temporary management where data is deleted after a specified period. They may also support embargoed data, which can be made public after a specified period of time.

Domain repositories focus on the needs of a specific field so they often have tools designed for data mining based on the research needs of their field. Domain repositories may only accept specific types and formats of data, not all areas of research currently have a corresponding domain repository. Domain repositories can also develop metadata to expose essential characteristics of their specialty, and management tools for the particular file types that researchers in their field employ. Domain repositories may be privately funded, public but funded by fees paid by researchers and/or publishers, or government funded.

Both types of repositories may provide public access to data under the conditions required by funding agencies. Institutional repositories also provide data storage for private archives, which most domain repositories do not support. A private archive is needed in cases where the data must be security stored, but can't be made available to the general public because of restrictions on privacy, copyright, trademarks, or patents. We will talk more about when you might choose to use a private archive and various options to preserve sensitive data next week.