Guidelines for Electronic Records Storage Systems

When setting up an electronic records storage system, you must ensure that the chosen system follows these essential guidelines.

Supports Multiple File Formats

Any system that you use to store your electronic records must support standard office file formats. At most institutions, however, you may need to ensure support for other file formats, including image files, and even audio and video files. Your system must be able to provide users with ready access to readable copies of the files stored in the system. Your system must also be capable of storing and providing access to any additional copies required for preservation, including preservation masters and the original files.

If at all possible, it is best to limit the number of formats retained within the system in order to reduce complexity. At the very least, a limited number of preservation formats must be used for the master files in the system. The best way to reduce the number of formats is by agreeing with the records creators on the formats that digital records will be transferred to you.

Provides Adequate Capacity

An important consideration for your storage system is ensuring you have enough space to store digital records. Your storage system must be able to accommodate future collection growth through such means as scalable storage configurations. The storage media you choose to use in your system (hard disk, magnetic tape, cloud storage, or optical media) depend on your available resources and the access needs of your users.

Supports Automated Processing

The system you use must be designed to provide an efficient way to move records into the storage system (a process archivists call "ingest"). The system must document any activities that take place or changes that are made to digital files as part of the ingest process. Typically, adding files to a preservation system requires a number of steps, many of which are best accomplished automatically:

Verifying file formats

Your system must verify the records are in the file formats you expected, that the files are not corrupted, and that they are still functioning. If the files are not what they appear to be, you will need to determine what they are before you can continue with ingest. If the files are corrupted in some way, you will have to determine if uncorrupted copies exist or if there is some way to fix the files before you continue. You need to do this work up front so that you are not surprised by problems with the files later on.

Checking for malicious code

An essential next step is to check all files for malicious code (computer viruses, Trojan horses, worms, spyware, etc.) before moving any files into your storage systems. You can do this simply by using an up-to-date malware detection program to scan the set of files. Some institutions wait an additional month afterwards, update their malware detection software, and run the scan again, just to be sure no newer and thus previously unrecognizable malware was stored in the files. Ensuring that malicious code does not enter your storage system is essential to protecting that system.

Converting to specific versions

In some cases, you may not be converting files at all, but simply ingesting files into a storage system as is. However, sometimes you need to convert a file to a preservation format for permanent storage, and you might also be creating a smaller use copy at the same time. See below for more on different versions of digital files.

Assigning metadata

As you prepare to move digital files into your storage system, you must assign metadata to the files to support the management, protection, and use of the records. One piece of essential metadata that can be assigned automatically upon ingest is a checksum, which is a unique number created by running an algorithm against a piece of digital data. Having this number allows you to verify, in the future, whether the file has later been corrupted. It is also good practice for the system to automatically create metadata on who added files to the system, added descriptive metadata, and performed preservation actions, and when they conducted this work.

Storing in assigned locations

At the point of ingest, you will also be storing files into logical and physical locations based on rules you have set up beforehand. See below for other details on considerations for storage for different versions of digital files.

Supports Adequate Metadata

Metadata is any information about the digital files in your storage system that helps you understand, manage, and provide access to those files. In an electronic records storage system, it is important to have enough metadata to support these activities well but not so much that the time spent creating and managing metadata exceeds the value it creates. Metadata created by your storage system should conform to accepted standards including, but not limited to, Describing Archives a Content Standard (DACS), Dublin Core, and Preservation Metadata: Implementation Strategies (PREMIS).

Descriptive metadata

Descriptive metadata provide information to help people search for and find a digital file. Common examples of descriptive metadata include titles, authors or creators, and subjects. A common form of descriptive metadata is simply the Library of Congress Subject Headings, which provides a controlled vocabulary for assigning subject terms to cataloged records. Your storage system must include descriptive metadata that can be integrated into your existing system of finding aids. Ensuring that the descriptive information provided by your storage system conforms to DACS, Dublin Core, or other accepted descriptive standards will facilitate this integration.

Administrative metadata

Administrative metadata consist of information needed for you to properly manage the digital files in your electronic records storage system. Such metadata include technical information (such as file formats, file version, and size), intellectual property rights data (such as copyright holder, owner of the file, and access restrictions required by the donor), and security metadata (such as physical or legal access restrictions). An essential part of your administrative metadata is a unique identifier that will link the digital object to its metadata.

Preservation metadata

A particular subset of administrative metadata is preservation metadata, or information that supports activities that ensure the long-term usability of digital materials. PREMIS is a widely accepted standard for preservation metadata, and it records information about

- characteristics of digital materials in your system, including fixity information
- actions taken to preserve the materials
- people involved in the management of the materials
- rights and permissions relevant to preserving digital materials

To best support good digital preservation, you should seek to have a digital storage system that incorporates PREMIS. For more information on PREMIS, see *Appendix 4: Relevant Metadata Standards*.

Structural metadata

Structural metadata is used to describe the relationship between a digital file and its component parts or how a set of digital files is organized and how the individual files relate to one another. Structural metadata is often used to support how digital files are stored and presented. For instance, if you are retaining a complex set of webpages interrelated to each other, then you would need to store them in such a way to support their intercommunication and proper presentation. Structural metadata is also used to describe complex databases and can document the relationships between tables in a relational database as well as provide information on the values of different data fields in the database. Metadata Encoding and Transmission Standard (METS) is a commonly used metadata standard that supports structural metadata, though it also supports descriptive and administrative metadata. For more information on METS, see *Appendix 4: Relevant Metadata Standards*.

Supports System Migration

The hardware and software you use for your system will become obsolete over time. For such a periodic system migration to take place as efficiently and error-free as possible, the original system must be open and standards-based. Any metadata within it must be stored using standard schema, and the earlier system should not contain other data or structural complications that make it difficult to carry out such a migration. See the next item for information on data conversion.

Supports Future Conversions

Since file formats will also become obsolete, your system must be able to support necessary file conversions in the future. A preservation file format that seems reasonable today may seem a bad bet in a few years, so you need a plan to continuously

assess the preservation formats you are using and to indicate the processes you will use to conduct large-scale conversions in the future.

Supports Storage for Different Versions of Files

In the digital world, you always must maintain more than one copy of a single record. At minimum, you need a master copy and a backup copy, but often you have other copies, each with different storage needs. Depending on the available technology and the needs of your users, you will need a system that a combination of online, nearline or offline storage. Regardless of the storage configuration of the system, it must be able to synchronize these various versions of items in your system.

Original

The original copy of a record is usually retained even if the file has been converted to a digital preservation format for long-term use. In cases such as these, the original can be retained in nearline or offline storage, since the need for access to it is minimal. Originals require storage that keeps the files from suffering bit rot and being affected by outside intrusions. Access to originals must be limited to IT and archives staff.

Master Copy

In some cases, the master copy and the original will be the same copy. This happens, for instance, when the original is not converted into a preservation format because it has been received as one. However, the master may also be a copy converted into a preservation format. The storage needed for master copies is the same as for originals, with access also limited to IT and archives staff.

Access Copy

The access copy, also called the use copy, is that copy that patrons are allowed to use. An access copy might be in a different format than a master copy, especially if that allows the use copy to take up less digital space. The use copy may also be redacted to keep sensitive information from being exposed, whereas a master copy must never be redacted. The use copy must always be stored separate from the master copy to decrease the chances that both will be destroyed by the same disaster. The access copy must be stored in a way that supports your ability to retrieve the record quickly, and it may be stored so that it is searchable and retrievable by patrons via an online or in-house system.

Backup Copy

For a properly functioning electronic records storage system, you need to ensure that all versions of your records (originals, masters, and use copies) are backed up and stored at a secure, remote location. These backup copies must be stored either on computer tape or in high-quality cloud storage. The tapes holding any backups must be refreshed for new tapes every three to five years. The location of any backup storage location must be geographically remote and chosen to ensure that the backup location will not simultaneously suffer the same disaster as the home institution.

Records Requiring Greater Security

For records that require high levels of security, you can maintain records accessible to only a select few on an access system, or you could keep such records in an offline dark archive that is rarely accessed. The latter provides more security since access can never be made of the records remotely.

Provides Adequate Security

You must always protect restricted or confidential information under your control, and you will do this initially by identifying the kinds of such information you hold. Many records in your institution will be confidential because they include personally identifiable information (PII). Such information is identified in a number of state and federal laws, including the Federal Education Records and Privacy Act (FERPA), and the Health Information Portability and Accountability Act (HIPAA). Other records may be restricted because donors have specifically limited access until a certain time in the future. Institutional records also may be identified as restricted because they contain confidential contract negotiations or sensitive research.

Once you know the kinds of restricted electronic records your institution holds and have identified those records, then you can begin to develop a system that provides adequate security. For the most sensitive materials, such security might require offline storage of those records. For general records security, you have to develop standards for both the hardware and software system's security and for the network that system is attached to. At the very least, this will include firewalls, continuous malware scans, daily intrusion tracking, and rules for staff use of these systems.

A most basic level of security is the use of account logins. Account logins, coupled with a requirement to use strong passwords, provide a basis for much of the access control of a system. With such controls in place, people without logins are barred from using the system. Logins also allow the system administrator to control people's ability to modify,

move, or delete files. Appropriate access controls also allow you to ensure that only a limited number of people have access to the most sensitive information.

For even greater security, your system could be set up to automatically create audit trails that record every action a user makes in the system. A system may include several audit trails, each devoted to a particular type of activity.

Finally, do not forget that the security of electronic records also requires that you store your digital records in a secure physical location as well.

Safeguards against Corruption

The system needs to have a verification program to ensure that the files stored within it have not been corrupted or changed in any way. One way to provide for this is through the periodic use of checksum software to monitor the contents of a digital archive for data loss or corruption. A checksum is a unique numerical sequence that is the outcome of running an algorithm on a piece of data. Comparing the checksum generated when a file was originally stored in your system with those created periodically afterwards, you can verify that your copy of the file has remained error free. In cases where the resulting checksum is not identical to the original, you will have to investigate the cause of the problem, solve it, and replace the corrupted file with an uncorrupted copy stored in a backup. Any corruption detected by your monitoring system must be documented by the metadata associated with the affected records.

Creates Backups at a Reasonable Frequency

You must decide how frequently to back up your data. Most systems are backed up entirely every day, and this is usually sufficient for your original, master, and access copies that do not change frequently. If you believe that your data requires more protection than this, you can consider creating a continuous backup to a mirrored site, but the cost of this will likely be prohibitive. Your most important role will be to ensure the backups are produced on schedule, verified for accuracy, and stored in a location remote from your central electronic record storage.

Minimizes Demand on Staff

To ensure efficiency, the system should make it easier for the staff to access and be able to complete their daily work. Whenever there are opportunities to automate tasks, especially repetitive ones, the system should allow it. Ideally, the system would also include templates for adding metadata to a series of records.

Relevant Metadata Standards

Dublin Core

The Dublin Core Metadata Initiative defines and provides information on Dublin Core in this way:

> The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. The name "Dublin" is due to its origin at a 1995 invitational workshop in Dublin, Ohio; "core" because its elements are broad and generic, usable for describing a wide range of resources.
>
> The fifteen-element "Dublin Core" described in this standard is part of a larger set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI). The full set of vocabularies, DCMI Metadata Terms [DCMI-TERMS], also includes sets of resource classes (including the DCMI Type Vocabulary [DCMI-TYPE]), vocabulary encoding schemes, and syntax encoding schemes. The terms in DCMI vocabularies are intended to be used in combination with terms from other, compatible vocabularies in the context of application profiles and on the basis of the DCMI Abstract Model [DCAM].

For more information on Dublin Core, visit the Dublin Core Metadata Initiative's website at http://dublincore.org/.

METS

The Library of Congress defines METS in this way: "The Metadata Encoding and Transmission Standard (METS) is a metadata standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of

the Library of Congress, and is being developed as an initiative of the Digital Library Federation." For more information on METS, visit the Library of Congress' website at http://www.loc.gov/standards/mets/.

PREMIS

The Library of Congress describes PREMIS in this way: "The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation."

For more information on PREMIS, visit the Library of Congress' website at http://www.loc.gov/standards/premis/.