Digitizing Archival Material Guidelines

from G.S.O., Box 459, Grand Central Station, New York, NY 10163

INTRODUCTION

We are living in a world of evolving technology. An increasing number of archives and libraries are using digital repositories and experimenting with technology, which may not always result in a desired outcome. Creating a digital repository can be a challenge, partly because it involves a variety of divergent subjects and issues. The methodology behind creating digital repositories, an awareness of the costs involved, and the ability to make decisions on what to digitize are all areas you might need to address when thinking about building a digital library.

Moreover, you will need to understand the technical criteria for the different materials in your collection, such as handwritten documents, photographs, magazines, books, audio, film, and more. The security of the repository, questions of access and use, as well as the issues of servers and back-up computer systems are also key points to consider.

As the General Service Office Archives regularly receives questions from local A.A. Archives on digitizing collections, we decided to address some of these issues. These guidelines are intended to be informative; they do not include everything there is to know about managing digital repositories, but they present a good start. No repository, no matter how large or small, will be able to digitize their entire collection. Digitization can be costly, time-consuming and some materials may not be fit for digitization or even worth the effort. Not all area or district archives can afford to digitize their collections, and that is okay.

MAINTAINING PRIVACY AND ANONYMITY

Protection of the privacy and anonymity of A.A. members is a critical issue that must be examined when maintaining digital files. The development of procedures seems to be a matter of local policy decision by the archives committees, but the necessity for protecting the confidentiality of correspondence and the anonymity of the correspondents is, without question, an important consideration and a trust that falls upon all A.A. archivists and archives committees. You may want to consider implementing a policy for using the digitized material and include guidelines for posting material on the Internet. For more information on anonymity online see the G.S.O.'s A.A. Guidelines on the Internet (MG-18).

WHAT IS DIGITIZATION?

Digitization is the process of converting information, such as text, photographs, audio and video files, into digital format. Digitization allows for the preservation of the content of the material by creating an accessible surrogate, putting less strain on the original. Digital objects are then maintained in a digital repository that offers a convenient way to store, manage, access and preserve these surrogates.

There are three basic digital repository functions. The first is the acquisition or capture of digital content. The second is the storage and management of digital content. The third entails the retrieval of digital content and creation of deliverables (or what can be done with the digital content). Hardware and software are needed for these functions. Their cost can range from zero to expensive based on the choice of hardware and software.

BASIC CONSIDERATIONS IN PLANNING YOUR PROJECT

Before establishing a digital repository consider some of the following questions: What are your reasons for developing a digital repository? Which reasons are most significant to your archives' mission statement? For example, some of the reasons for digitizing might include increased access and use, security, preservation, management and authenticity. At the General Service Office Archives our digitization efforts have aided in improved access, as well as improved preservation of the materials by reducing handling of the originals. You will want to take time to decide upon which digitization equipment would be best for your repository. Basic research should be done prior to purchasing hardware and software. Look into possible threats to your repository, including media, hardware, software and/or power failures. Here are some questions to consider that may help you to avoid disastrous consequences:

Would your archives be able to afford long-term financial sustainability of the hardware and software needed?

- On what media type are you storing the data?

- How would your storage devices be backed-up?
- Have you considered hardware and software obsolescence?
- How would you control access to the data?
- How would you protect the data from electronic malware, virus, etc.?

Software requirements:

First, decide on the type of scanning application would best suit your needs. For small, local collections, simple Windows or MAC based applications may be used. For a significant collection containing at least 5,000 or more images Content Management (and file management) software is recommended.

It is important to have good software tools for scanning, file editing and PDF file functionality. Many scanner vendors bundle software with the driver software needed for the scanning device. There is an abundance of open-source software that can be downloaded from the Internet. Many software and hardware vendors offer reduced prices to nonprofit entities.

If a program is open-source, its source code is freely available to its users. There are no licensing fees or other restrictions on the software. Thus, a user has the ability to take this source code, modify it and redistribute it. Open source software rarely comes with technical support and users rely on an online community for guidance and support. Technical expertise is needed to use open-source software. On the other hand, proprietary or closed-source software consists of programs distributed by a trusted brand for purchase. Under the licensing agreements for proprietary software are user restrictions that ban modification and redistribution. Proprietary software offers user support, which is typically an attractive quality for users without advanced technical skills. Regardless of which type you use, it will require some modifications and someone in the IT profession to get your system up and running. It is important to carefully review the Terms and Conditions of your software license prior to any signed agreement. One important consideration to address is to know what levels of control are maintained by you with regard to accessing and editing your data, should your software license expire.

Hardware requirements:

Basic hardware needs will typically consist of a personal computer (preferably one with a high processing speed), a high quality monitor, a CD-ROM drive, a 40GB (or larger) hard drive, a laser color/black and white printer, a scanner that can potentially support photographs, and an optional hand-held scanner. Determining what material is to be scanned (text, photographs or artwork) impacts your decision on what equipment and the size of the scanner you should select. Flat-bed scanners are recommended for most digitization projects, while a hand-held scanner is useful for oversized material, fragile material, and bound-volumes.

Storage media for the repository's master files is probably best satisfied by hard-drive systems. It is important to note that storing files for access to CD-R, DVD-R, Blu-Ray, and particularly flash-drives will not guarantee longevity of your master files. However, if material is stored on CD-Rs or DVD-Rs, it is recommended to use high quality or "archival" quality CD-Rs and DVD-Rs (such as Mitsui Gold Archive CD-Rs) and to store these discs in a temperature and humidity controlled environment.

Software and hardware technologies require ongoing attention due to continuous and rapid advancements. When a new technology emerges it usually quickly replaces its older version. When a software technology is abandoned or a hardware device is no longer produced, digital records created with such technologies are at risk of loss. This is called digital obsolescence. Oddly, today we can still read centuries-old original documents in their native language. Yet, in fairly short order, digital media technology has advanced from floppy disks, to diskettes, to CDs, to DVDs, to Blu-Ray disks and flash memory media. Remember that any storage medium can fail at any time. Archive your data on more than one medium and check your archives regularly for failures.

Be aware of "Cloud Companies" that may appear to be an ideal data backup storage option. Cloud backup involves sending a copy of the data over the internet to an off-site server. The server is usually a third-party service provider, who chargers a fee based on bandwidth, capacity, or number of users. Most of these companies assume complete control on the data you have deposited into their custody. Before you leap into the cloud, examine carefully the Terms of Service and consider these significant issues:

- Security Issues. How safely is your data protected?
- What happens if you decide to terminate your service? Can the cloud company keep a copy of your data?

- What happens if the company goes out of business and sells your data?
- Inflexibility. Be careful that you have not locked your data into a proprietary application or format.

SELECTING ORIGINALS TO DIGITIZE

Many archivists are faced with the daunting question of what to digitize. Some learn about the digitization efforts of other repositories and think that they are just not doing the right thing. Rest assured that your efforts in preserving A.A. history are a good start. We cannot tell you specifically what to digitize because the scope of your collection is unique.

Be selective in the choice of planned digital content by focusing on quality not quantity. We also recommend that you focus on digitizing material that pertain to your area or district, such as minutes, fliers, correspondence, and other items generated by your local A.A. entities. Digitizing these items will provide you with an electronic duplicate, which is easy to access and use, yet sparing the original item from overuse.

Some materials are born digital, meaning they originate in a digital form (e.g. desk-top publishing, digital cameras, etc.). However, they should be imported into your digital repository and saved in a way that adheres to the file name scheme already agreed upon for digital objects. Textual documents can be digitized if no harm is done to the document during the process.

Documents that do not fit onto a standard flatbed scanner are called oversized documents. These documents, as well as bound books, should be digitized using a hand-held scanner, scanning digital camera, or a standard digital camera. It is not recommended to scan bound volumes on a flatbed scanner as this may cause permanent damage to the book spine and binding.

BUILDING A STRUCUTRE FOR YOUR DATA

Now that you have adequately addressed these significant questions, formulate a plan on organizing the data in some sensible order. This is also a time to decide on a file name convention for your folders and files. A file naming system should be agreed upon prior to digitizing in order to create consistency. Descriptive file names should relate to the item being digitized and may contain the name of the item or accession number. You might consider modeling the digital file name system based on the paper file name system used in your repository.

The Archives Workbook contains a page titled a "Calendar of Holdings" for G.S.O. Archives. It's a high level structure of how the G.S.O. Archives' collection is organized and categorized. This same type of simple structure should be used to organize the content of digital repositories.

The following is an example of the folder structure in the G.S.O. Archives electronic database:

Newsletters (Parent Folder)

- ➦ *About A.A.* (subfolder)
  - *About AA*, 1972 Summer (file name)
- ➦ *Box 4-5-9* (subfolder)
  - *Box 4-5-9,* 2010 Fall (file name)

METADATA

Any individual engaged in a scanning project will come across the term "metadata." Simply put, metadata is "information about information." It is information used to describe, locate and retrieve files stored in a digital library. It is the key to ensuring that electronic resources will continue to be easily accessible in the future. In Microsoft Windows and other software, metadata may also be called "properties." The two words mean the same thing.

Most metadata, such as file size and date, is automatically generated when a file is created. Application software also generates metadata when a file is created, such as the file name and type.

Names and values should be simple and consistent. One way to achieve this is to use a controlled vocabulary. This is simply an organized list of words and phrases used to tag digital content and retrieve it through search. Use unambiguous and meaningful folder and file names and descriptive keywords. A file name is metadata and should offer a description of its content.

In the library environment there are several complex standards or schemas that exist for describing digital files for different types of material (books, photographs, audio, etc.), and each will have its own unique structure. However, as long as consistency is maintained, you can easily create your own simple metadata schema to fit your repository's resources and needs.

The software program your repository chooses to use will have its own mechanism for adding metadata. For example, the G.S.O. Archives electronic repository

contains thousands of group documents. In order to find information for just the right group, we built a metadata structure that is unique for only the Group records. The fields include: “Name of Group,” “Group Number,” “Meeting Place,” “Area,” “District,” “City,” “State” and “Country.” This allows the program to efficiently locate a group by inputting data in any one of the fields. There is more to building metadata than is described above, but remember: keep it simple and maintain consistency.

DIGITIZATION SPECIFICATIONS

A variety of factors will affect the appearance of images, whether displayed or printed. Therefore, when scanning, there are different resolution and file type requirements for different types, formats and the character of the originals. However, there are far too many unique document characteristics, each requiring a recommended parameter.

- When scanning documents with printed type (e.g. laser printed or typeset) it is generally recommended to use a minimum of 300 ppi at a 8-bit grayscale mode.
- If scanning documents with poor legibility, handwritten, carbon copies or photographs, the recommended parameter is 400 ppi at 8-bit grayscale mode.
- Documents requiring an accurate representation of the original color should be scanned at 300- 400 ppi with a 24-bit RGB mode.
- Open file formats should be used as they are publically obtainable and are recognized by multiple programs. The table below provides for the most common types of open file formats used in digitization project.

Open Office XML is also an open file format used in Microsoft Office 2007, 2010 and 2013. Examples of these files include DOCX (word processing document), XLSX (spreadsheets) and PPTX (presentations). Older versions of Microsoft Office may not be in an open file format and therefore can only be opened using Microsoft Office.

QUALITY CONTROL

A quality review of the scanned images is a highly critical and significant aspect of the digitization process. It is also recommended that regular inspection of the scanner be performed to ensure that physical matter, such as dust or specks of paper, is not left on the flatbed.

Remember to always keep your work area clean, which includes daily maintenance of the scanning equipment. Many old documents or books tend to be dusty or even moldy and will leave potentially harmful substances on the flatbed. Visual checks should be done periodically to the digital image files to ensure quality is maintained in the following areas:

- The correct format (TIFF, JPEG, etc) and resolution are used.
- The digital image has the correct orientation.
- For multi-page documents, all pages accounted for and in correct order.
- The digital image is complete and not cropped.
- The image is not skewed.
- Digital artifacts (dust, specks of paper, etc.) do not appear on the image.

File Format	Technical Characteristics	Recommended Use
TIFF (Tagged Image File Format)	• Supports most platforms. • Lossless (there is no quality loss due to compression). • Accommodates large file sizes. • Flexible format. • Preferred image format for preservation.	Recommended format for archiving. It is most universal and widely accepted.
JPEG (Joint Photographic Expert Group)	• Acceptable for photographic images files if least compressed setting is used. • Quality is lost every time the JPEG file is compressed and saved.	Access file use only. Not recommended for production master files.
PDF (Portable Document Format)	• Complex file format.	Access file use only. Not recommended for production master file or for archiving. Can be processed for OCR text recognition to make file searchable.

- Image quality is maintained for color, tone, sharpness and contrast.
- Metadata related — the digital files are named properly.

PRESERVATION OF ORIGINAL MATERIALS AFTER SCANNING

The end result of a digitization project is a digital surrogate of the original. It is important to take proper care of the original material. Although you will have a digital copy, it is still vital to protect historical documents from deterioration by using archival quality material and optimum environmental conditions for storage.

See the G.S.O. Archives' Preservation Guidelines for a more comprehensive and in-depth look into the preservation of archival material.

PLEASE NOTE...

For answers to any specific questions, and lists of additional resources, feel free to contact the G.S.O. Archives at archives@aa.org or 212-870-3400. Other valuable information is available on G.S.O.'s A.A. Web site, www.aa.org.

SOURCES FOR MORE INFORMATION:

Please note that the G.S.O. Archives does not endorse nor affirm these Web sites, and simply provided them as helpful external resources.

National Archives and Records Administration (NARA)
http://www.archives.gov/preservation/technical/guidelines.html

Digital Preservation Coalition
http://www.dpconline.org/advice/preservationhandbook