Issues in Managing Digital Archive Collections

Kellie Johnson

Emporia State University

LI 855XU

Abstract

When dealing with archival collections in any format access is always the key to a viable repository. At this point in history researchers are practicing the unique ability to travel to archives the world over with a few taps of the computer keyboard. This paper will cover three main issues in the management of digital archive collections culminating in a fourth and final topic of access. The proper care and management of a digital archival collection through longevity management, integrity, and data loss security will ultimately allow for greater access.

Issues in Managing Digital Archive Collections

In the twenty-first century the traditional musty archive has gone high tech. No longer need researchers trek to far off destinations to find primary source information when many institutions are making most everything available to anyone with access to a computer. While establishing a digital archive collection may seem to be the most difficult part of the equation it is actually the ongoing management of that very same collection which will prove to be the most complex. This paper will discuss briefly four areas of concern including: longevity management, integrity, data loss security, and access.

Longevity Management

Every year things change in the computer and information technology industry. Computer hardware, software, and the electronic gadgets to access information on, change so quickly that even television commercials spoof the anxiety of an out of date public. This process is also felt in the archival world. Claims of permanence from manufacturers have not stood the test of even a small period of time. According to Kevin Bradley's article *Defining Digital Sustainability* (2007), "The goal of a permanent media has been wrecked on the rocks of relentless progress" (p. 153). He goes on to note that even if the software or storage schemes were permanent the items used to access the information would no longer exist or become unviable do to attractiveness of speedier systems (Brandley, 2007). As a result it is best for archivists to avoid any kind of proprietary software to ensure fewer difficulties in the future.

Still, archivists are left with the knowledge that at some point some original digital format files sitting on everything from magnetic tape to CDs will be unreadable at some point in the future. While some are content with keeping around old computers with five inch floppy drives others are looking into solutions to file reformatting. These solutions include conversion, migration, and emulation.

Conversion

Conversion of digital files is a process most people are familiar with. This happens quite often when someone emails a file that they created using WordPerfect. However the receiver has only Microsoft Word. What is good about this particular situation is that Microsoft Word can open the WordPerfect file by converting the format to fit its particular software system. Format is not lost, however some underlying data may be altered (Hunter, 2003). This process is extended to various formats of digital archives in which workers periodically will convert files to work with the latest in hardware devices and software packages. While the overall information may not be lost there is a sense that the integrity of the file has been changed.

Migration

Migration of digital files, according to Frederick Stielow (2003), is very similar to that of conversion involving the periodic recopying of storage formats (diskette to CD), hard drives, software (WordPerfect to Microsoft Word), and languages (HTML to XHTML). However, Hunter (2003) describes migration as a "last-gasp attempt to move digital information from long-inactive legacy computer systems" (p. 265). This happens more so than archivists like to admit.

Within the U.S. national government backlogs of information cause formats to pass out of existence before items can be properly added to the digital record. Engineers are employed to write migration software programs to transfer materials from one format to another. Sometimes digital archeology is also employed to assist migration where hardware needs to be refurbished and/or constructed due to obsolescence. Recently Dennis Wingo of the Lunar Orbiter Image Recovery Project (LOIRP) has been able to transfer NASA Lunar Orbiter magnetic tapes into digital images by reconstructing FR-900 tape drives that were stored in a barn for several years (National Aeronautics and Space Administration, n.d.; Pringle, 2010). This project ultimately combined both digital archeology and software migration via magnetic tape to digital formatting. Here again underlying data can be lost whether through software or hardware transfer.

Emulation

Due to concerns of data loss Jeff Rothenberg proposed, in his often cited piece *Avoiding Technological Quicksand*, the concept of emulation. According to Rothenberg both practices of conversion and migration take too much time and allow for data loss. Emulation on the other hand is the "only reliable way to recreate a digital document's original functionality, look, and feel" (Rothenberg, 1998). This is done by embedding the original digital file with a program that will allow the reproduction or emulation of that same file on future computer systems.

The program according to Rothenberg (1998) should be encapsulated within the original file format specifying emulators to work on future computers and saving metadata in human readable formats for accessibility and data recreation. While this is still hypothesis Creative Archiving at Michigan & Leeds: Emulating the Old on the New (CAMiLEON) is working to make emulation a viable tool in digital preservation. According to their official website CAMiLEON has had some success with the emulation of the BBC Doomsday laserdisc, however no official finding reports can be found at this time (http://www2.si.umich.edu/CAMILEON/).

At the same time as the CAMiLEON project Raymond Lorie was working on the similar emulation concept of Universal Virtual Computer (UVC) in which, "The main idea consists of archiving a program P along with the data file that decodes the data and returns the information to a future client based on a logical view . . . without any specific software or hardware" (Lorie, 2002, p. v). However, good the idea of UVC may be it has not seemed to have gathered any traction within the archival field (Bradley, 2007). For all emulation promises for preservationists, in whatever format it appears, the everyday application may be some years off still and archives will more than likely depend on aspects of conversion or migration as the needs arise. Still, the need of some sort of process is necessary to ensure continued access to digital archival files.

Integrity

Digital archive collections come in two basic formats. The first format is that of primary sources materials that have been digitally scanned as to allow for greater access to content. And the second format is that of items which are born digital. These items can include everything from electronic mail to web postings to this very paper. However, the potential for manipulation of the digital file can leave archivists and researchers alike in a position of having information that may be invalid. With this in mind, this section will deal with reliability and authenticity of digital files along with metadata solutions.

Reliability and Authenticity

Both reliability and authenticity deal with the integrity of the digital file. First reliability is defined as referring to "the authority and trustworthiness of a record as evidence of what it is about, that is, to its ability to stand for the fact it speaks of" (Duranti, 2002, p. 25). This means that when compared with other known data sets, the information contained in the first set is reliable as to the known record. The information should never be changed however it is not so hard to presume that some individuals will invent information so as to draw others into false or misleading conclusions.

Image files are some of the best known examples of questions of reliability in records. Magazines are notorious for altering digital files to create images that do not exist in reality. One such recent example is the February 7, 2011 cover of *Time* depicting a very much alive former

President Ronald Reagan with his arm around current President Barak Obama (Time, 2011). While some do know that Reagan has died future generations may not be so sure unless they are able to verify the reliability or lack thereof contained in the *Time* cover image. While reliability is the responsibility of the record creator it is the archivist's duty to incorporate security measures, making sure that access to files does not allow for manipulation of the known record.

Secondly, authenticity deals with file integrity in that it presents a provenance of the record from creation through the latest migration. Duranti (2002) states that, "A document is authentic if it can be demonstrated that it is precisely as it was first transmitted or set aside for preservation" (p. 27). Some have brought up issues regarding what is actually authentic when dealing with digital preservation. According to David Ryan (2007) these matters can be compared to the legality of the file noting that several institutions around the world track the authenticity of electronic records. Was the record created by whom and when it claims to be? This question can be answered with the maintenance of reliable metadata.

Metadata Solutions

Metadata is basically data about data. Or in this particular case the data about the digital archive file. Essentially the metadata allows the archivist to track information throughout the digital file lifecycle. This is attached to the file representing information relating to creation, migration, and even disposal. Provenance data including when, where, and by whom the file was created is essential information relating to the authenticity of the file. Format of the original file creation as well as basic descriptive information about the contents of the file should also be included.

Accordingly, it may be necessary to migrate or convert the digital archive at various points in its lifecycle. This information should, at that time, be added to the metadata. Archivists should specifically include information involving matters of file format change, e.g. from file format A to file format B. This process will cause metadata to accrue over time and systems will have to be large enough to handle all of that information.

As files go through the lifecycle it may be necessary to dispose of items for various reasons. However, the metadata file may be the only record of a particular item and what has happened to that item. This is why Kate Cumming (2007) suggests that metadata may need to be kept longer than the actual digital file. Some institutions, due to legalities, may have to have a permanent record of all items including "authorizations of disposal" (Cumming, 2007, p. 48). It is a good practice for any archive to institute this particular process, whether or not it is mandatory. Archivists should be aware of metadata management which confirms the reliability and authenticity of any digital archival file and know that it is an essential part of the ongoing management that same file.

Data Loss Security

Digital archival files are at risk to issues of security just as much as their paper counterparts. Some concerns relate to matters of integrity and keeping the record in an as authentic state as possible through ongoing processes of conversion, migration, or emulation. However, digital archives are also prone to everyday problems such as physical file storage media being affected by degradation over time or elemental disaster; while computer files are prone to data loss due to viruses, system crashes, and electrical power outages.

Archival science has adopted the concept of backing up both physical and virtual systems in multiple formats. Many institutions may carry a copy of files at a separate physical location to ensure safety if anything untoward might happen to the original object or file. This section will concentrate in particular on digital archival data loss. Three areas will be covered the first being a prominent system developed for digital archives called Lots of Copies Keeps Stuff Safe (LOCKSS). Second will be the use Computer Output Microfilm to back up digital copies, while the third option to be discussed is cloud computing.

Lots of Copies Keeps Stuff Safe (LOCKSS)

Lots of Copies Keeps Stuff Safe (LOCKSS) has become a mantra in the archival field whether it pertains to establishing a system of physical copies of items or digital files. As it pertains to digital files LOCKSS, based at Stanford University Libraries, allows members to cache data that will be accessible via institutional accounts when system failures arise, including server crashes, according to Reich and Rosenthal (2009), as well as the official LOCKSS website (http://lockss.stanford.edu/lockss/Home). According to the website LOCKSS also allows for access to files inputted into its systems even after cancelation of services as well as performing automatic migration services on the fly as access requires.

The LOCKSS website also includes a quote by Thomas Jefferson, ". . . let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident." LOCKSS ultimately allows access to information no matter what problems the main institution may be experiencing. While this is a program that requires service and participation fees in order to access content, the concept of keeping multiple copies has caught on with archival institutions.

Computer Output Microfilm (COM)

Computer Output Microfilm or COM is exactly what it sounds like. This procedure involves taking digitally created files, born digital or digitized items and printing them onto microfilm. In this instance metadata will be lost through the transformation from digital to microfilm unless it is part of the content of that file. While this may not be a best practice for all institutions, government organizations that deal with great numbers of digital files may consider this a viable alternative to printing out massive volumes of information. Both the New York State Archives (Gavitt, 2002) and Utah State Archives (2008) websites note considerations as to cost, space, access, and archiving as major draws to the use of COM. Microfilming allows for secure storage of information that can be placed at localities distant from where the original digital files are stored.

Cloud Computing

Secure storage of digital information in locations other than an institution's physical location can be aided by the forming trend of cloud computing. The concept is that individuals, business, and institutions alike can use servers located around the world as a back up to local hard drives and mainframes. Generally, clouding companies allow for security protocols which will allow for secure protection of digital archival files. However, storage in the cloud can have some drawbacks.

According to a 2010 article by Stuart and Bromage there are some problems when it comes to current cloud storage capabilities. Major areas of concern for the authors included authenticity and access. Some of these issues could be dealt with by:

- Asking providers where digital records will be stored
- Contracting for privacy agreements
- Asking providers to eliminate all records at contract termination
- Asking providers for information on back up storage (Stuart & Bromage, 2010)

This being said cloud computing may prove to be a viable solution for some institutions in the near future allowing for increased data loss security and research access.

Access

Ultimately the principle behind an archive is to allow access to information pertaining to everything from business to educational to governmental matters. While it is duly noted that some information remains inaccessible to certain persons whether due to privacy, legal, or security issues much information is accessible to the general public for scholarly or personal research. The most pertinent topics that can hinder researchers include those covered above including file longevity management, file integrity, and data loss security.

However, there are some more issues that researchers are concerned with in accessing digital archives. Alexander Maxwell's 2010 article on end-user preferences discusses some matters concerning digital archives involving history researchers. One particular annoyance, for Maxwell (2010), was the necessity to download proprietary software prior to being able to view any materials when using DjVu and DLibra. While he noted the software did allow for ease of maneuverability, which was another major concern of his, it also encumbered access for those who may not be able to download necessary software.

Maxwell (2010) also was displeased that the integrity of some historical information was sacrificed in order to better serve access. While some items were transcribed and made available via HTML it is more preferable to the historian, according to Maxwell (2010), to view an original image in PDF formatting due to inevitable and unforeseen human error. Articles such as these can give archivists the tools they need to create better, more accessible archives in the future.

Conclusion

While longevity management, integrity, and data loss security are of major importance to the management of digital archive collections they all deal with the central theme of access.

Without proper file formatting no one would be able to view digital files beyond their original inception. Integrity of the file allows for the authenticity and the ability to record the lifecycle of that same file. And data loss security not only allows for the protection of the digital archive, through diversification of storage capabilities, it also provides access through varied formats and systems. Access is the key to twenty-first century technology. Well managed digital archive collections need to be on the forefront of innovations pertaining to global access. Therefore the researcher will no longer be forced to go to the information, but the information will come to the researcher.

References

Bradley, K. (2007). Defining digital sustainability. *Library Trends, (56)*1, 148-163.

Cumming, K. (2007). Metadata matters. In J. McLeod & C. Hare (Eds.), *Managing Electronic Records* (pp.34-49). London: Facet Publishing.

Duranti, L. (2002). The reliability and authenticity of electronic records. In T. Eastwood (Ed.), *Preservation of the integrity of electronic records* (pp. 23-30). Boston: Kulwar Academic Publishers.

Gavitt, S. (2002). Computer output microfilm (COM). New York State Archives (Publication Number 52). Retrieved from http://www.archives.nysed.gov/a/records/mr_pub52.pdf

Hunter, G. S. (2003). *Developing and maintaining practical archives: A how-to-do-it manual* (2nd ed., No. 122). New York: Neal-Schumann Publishers Inc.

Lorie, R. (2002). The UVC: A method for preserving digital documents – proof of concept. Netherlands: IBM.

Maxwell, A. (2010). Digital archives and history research: feedback from an end-user. *Library Review, (59)*1, 24-39. doi: 10.1108/00242531011014664

National Aeronautic and Space Administration (NASA). (n.d*.). Lunar Orbiter Image Recovery Project (LOIRP) Overview*. Retrieved from http://www.nasa.gov/topics/moonmars/features/LOIRP/

Pringle, H. (2010). NASA dives into its past to retrieve vintage satellite data. *Science, (327)*5971, 1322-1323. doi: 10.1126/science.327.5971.1322

Reich, V., & Rosenthal, D. (2009). Distributed digital preservation: Private LOCKSS networks as business, social, and technical frameworks. *Library Trends, (57)*3, 461-475. doi: 10.1353/lib.0.0047

Rothenberg, J. (1998). *Avoiding technological quicksand: Finding a viable technical foundation for digital preservation*. Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/rothenberg/contents.html

Ryan, D. (2007). Digital preservation – 'the beautiful promise'. In J. McLeod & C. Hare (Eds.), *Managing Electronic Records* (pp.50-62). London: Facet Publishing.

Stielow, F. (2003). *Building digital archives, descriptions, and displays: A how-to-do-it manual* (No. 116). New York: Neal-Schumann Publishers Inc.

Stuart, K., & Bromage, D. (2010). Current state of play: records management and the cloud. *Records Management Journal*, (20)2, 217-225. doi: 10.1108/09565691011064340

Time. (2011). *Why Obama [heart] Reagan*. Retrieved from http://www.time.com/time/covers/0,16641,20110207,00.html

Utah State Archives. (2008). *Deciding to microfilm*. Retrieved from http://archives.utah.gov/micrographics/micrographics-guide-deciding-to-microfilm.html#COM