Forecasting the Future of Digitization Permanence through File Formats

Jessie Gillan
LIS 632: Conservation & Preservation: Final Report
April 25, 2009

"You may delay, but time will not." – Benjamin Franklin

The culture of today is creating and digesting more information content than was really ever expected for a person to be bombarded with. Workplaces have mostly thrown out all of their typewriters and create all of their important documents on computers; hand-drawn typography and graphics are now created with Photoshop. The world has shifted from a print world to a digital world, and librarians are under pressure to determine a way to ensure that the electronic data will not be lost or forgotten. The International Data Corp has estimated that "the amount of information that is created, captured or replicated in digital form in 2011 will be 10 times greater than that produced in 2006" (Levi 2008, 22). Many cultural heritage institutions around the world are beginning to embrace digitization as a method for allowing greater access to their materials. The greater access is a double-edged sword, on one hand it promotes the library and creates more interest in the library as a whole; on the other side, there are copyright issues and concerns that actual user numbers of the library will decrease with content being publicly available.

Librarians seem to be moving forward with digitization with a bit of hesitation in their step because of the digital medium itself. Digital content has only existed for the last 30 years or so, and the technology has changed so quickly throughout that span that it is hard to be sure that the current method will be the best choice. Every aspect of the technology has changed drastically since its inception, which has led to the librarian or archivist to try and forecast the future of the digital world. This essay will explore the past file formats to begin to adopt a

stronger sense of security towards the future of digital content as we strive for digital permanence.

Early History of Personal Computers

The first personal computer the Apple II was introduced in 1977 (Weyhrich, 2009). This was the beginning of "born digital" documents, which is the information or files that are created in an electronic format, commonly the personal computer. The storage abilities and overall quality of the earliest systems was very low and as such the transition to the personal computer from typewriter was very slow. Many errors were made in the preservation of many of the files and because of bad technological transitions between systems through the late '70s to the early '90s. The file storage sizes were quite small internally and for portable storage the floppy disks were fairly successful but were constantly being upgraded to a smaller model, which left the older larger disks to be unusable.

The components of all of this technology took some time to get together and work together well. The compact disc first came onto the market in 1982 (research for the disc began in 1969) and for a long time there was no cooperation by the computer companies to implement an optical disc drive into their systems to use compact disc technology (BBC News, 2007). Magnetic tapes are still seen as the best way to store and preserve digital records, so perhaps the compact disc and magnetic floppy disk war was the first attempt to halt the development of bad preservation tactics. Compact discs have had issues with quality from their onset and it is important when evaluating these materials today to make sure you are purchasing the best quality discs. If you are writing/burning compact discs in your institution please use the advice given in IASA-TC 04, "Guidelines on the Production and Preservation of

Digital Audio Objects" by Kevin Bradley. Compact discs have never been a successful long lasting method to store materials; they are easily scratched, worn away, and generally damaged. Using magnetic tapes to store materials has always been the best and most sound method to maintain digital content.

Additionally, Many of the first software companies initially made proprietary software that led to many different file types existing and an overall lack of consideration for the future of the information created. This mistake lead to many documents deleted or thrown away because they were developed on DOS proprietary software such as, WordStar, Lotus 1-2-3, and dBase. With all of these initial incorrect decisions much of the born digital information created during the early period of personal computers has been lost almost completely. Most of these file types can be retrieved and transferred to today's file extensions today, but it is already too late for most of these documents. Generally when a new system became available to an institution all of the documents were not transferred to the next system and floppy disks because of frequent size and type changes were thrown to the garbage.

File Formats: Past, Present, and Future

The most common file formats used within cultural heritage institutions today are PDF, DOC, and JPEG/TIFF (Levi 2008, 22). These file types are also the most popular with home computer users, which is an immediate sign that some standards have developed in the computer industry since the early days. To discuss the progress of these file types will inherently prove that our digitization methods today will save us from a digital file meltdown. Actually, it is clear we are heading closer and closer to a digital permanence.

The PDF (Portable Document Format) has become the most popular of all the file types because it is a great all-purpose format that can both hold images and text. As Judith Rog writes in "Recommendations for the creation of PDF files for long-term preservation and access," "Adobe Systems invented PDF technology in the early 1990s to smooth the process of moving text and graphics from publishers to printing-presses and has been in use since 1993" (2). Initially PDF files required Adobe Acrobat to access the content of the file, once the popularity of the format caught on they released Adobe Reader a free program that is open to the public without cost. Now, PDFs can be opened through some of the most popular browsers without having the Adobe software installed and also through basic image preview programs on both Mac and PC.

In 2005 the standard for PDF/A-1 was established as ISO 19005-1 (Turro 2008, 25). Establishing this standard was a great step in the long-term preservation of PDF files. Within PDF files there is a wide range of possibilities with metadata and tagging the content held within for catalog searching. Furthermore, documents scanned by into PDF using OCR (Optical Character Recognition) technology can be completely searchable by all of the words within their content. However, there are errors that occur in this process as well and not all words get entered correctly. A technology genius, Luis von Ahn, has created a system called "Re-Captcha" that uses words that OCR does not recognize for security text on website registration pages. This tool and his other creation, Games with a Purpose, people are assisting to create a more accurate digital heritage for the future (PBS Wired Science 2007). The PDF is a great tool for merging many files into a format that is easy to read, search, and access for many users around the globe.

The DOC file format is the word processing document created by Microsoft Word. This program has been used in offices, libraries, and homes for the past fifteen to twenty years and has become a standard because of its popularity. Recently with the latest update of the Microsoft Office suite they made the first change to their file format since the inception, adding an "x" to the end of their file names displaying that they now use a combination of XML. Microsoft has made a promise (Microsoft Open Specification Promise) that these files will be able to be more of an open file extension and will be able to be used on previous version of Office, however this has not happened and people are ignoring the change and saving their documents to the DOC format to maintain consistency.

Microsoft has had a wave of troubles in recent years with their programs and operating systems not working well when released to the public. The new Office seems to be enigmatic of those troubles. Many have switched to doing their typical Microsoft Office work through OpenOffice, available at OpenOffice.org, which is a free software package that works similarly to the Office Suite. OpenOffice creates open files that can be converted easily to any file extension or type. The future is a bit uncertain with DOCX files as of now, the newest Office does work with previous DOC files, which at least shows some foresight for the future stability of those documents.

A bit of a divergence between the digital preservation standard and the common use is in the image file formats JPEG (common) and TIFF (preservation standard). The TIFF (Tagged Image File Format) is preferred because of the ability to store metadata within the object through the TIFF file. The JPEG (Joint Photographers Export Group) is still used in preservation for a quicker loading compressed small file. These two formats are pretty widely accessible as

well, the best results can be created through the Adobe suite of Design products but the ability to do great work without these is also possible. GIF (Graphic Interchange Format) is the next most popular image format, and should not be used in digital preservation because as the SAA glossary states, "the image uses lossless compression to minimize file size" and has a limit of 256 colors. The commonality in the image file formats is really grand and quite often it is easy by changing the file extension to change the file to be readable by any system.

Steps to Creating Longer Lasting Digital Files

The first important note on files is to never encrypt your files with password protection. Experts can break this, but it is best to not let it get to this point. Keep private files on password protected computer systems; the files themselves at that point no longer need password protection.

Along the same line of thought, using "read-only" should be used wisely. A read-only file is compressed the file and locked it from editing. Publicly files should be marked as read only or made un-editable in another fashion. Library held files should generally not be read only so they can be updated when they need editing. Also, "read-only" could be a great function to maintain important letters from being wiped away and other important digital documents held in an archive, so it could be a function that saves your computer from a massive human error.

If you are placing a link in the document or image make sure you use the entire web address include the "http://" this will at least provide more information on how to re-link the page if the page moves or is deleted from the outside source. You can never be sure that a link will exist one day to the next so it is important to provide yourself with breadcrumbs to perhaps

find the document again. Also in this instance it could be good to use the Wayback Machine held at the Internet Archive to seek the information that they may have stored in their webpage archive.

In PDFs and DOCs it is important to use fonts that are in the "base 14 fonts: Courier, Courier Bold, Courier Bold Italic, Courier Italic, Helvetica Bold, Helvetica Bold Italic, Helvetica Italic, Times, Times Bold, Times Bold Italic, Times Italic, Symbol and Zapf Dingbat" (Rog 2007, 5). These fonts are universal on both MACs and PCs and have been in the default font sets since almost the onset of personal computers. If the document uses a different font than these you can either switch the font or choose to embed the font file in the document so when someone else opens the file the font is downloaded into their font library automatically. Also under style guides of the files it is important to use either RGB or CMYK color for image or PDF files could cause corruption down the line and loss of information. Every computer stores color profiles into image files and PDFs and these are maintained through saving and creating the files using CMYK or RGB color interfaces instead of the other available options.

Moving further on, the next key is to always place as much metadata/tagging into files as possible so both the users and librarians can find them in a search. Additionally, creating obvious folder and file names within your system helps others to find the information, never only use a number or other non-informational name for a file or folder.

In addition, keeping as many files in as many formats as possible is always a great way to ensure that one of them will work in the future. When you save your files you should always choose the option with the least compression and also save at the highest quality possible so the least amount of data is lost in the compression process. Every time a file is saved it loses a

small slice of the original data that was attached to it. If you must compress a digital file use "compression algorithms such as ZIP" (Rog 2007, 6).

Continuing further, never use scripting or other executable actions within your documents this relies on another program to maintain their function and could cause the entire document to malfunction in the future. Examples of scripting or executable actions would be using JavaScript in a PDF file or a QuickTime Movie file within a PowerPoint. These could cause further troubles with the documents and will not be great for long-term preservation.

Important Fixtures in Digitization

As discovered in the above, file types are becoming more and more standardized and such it is easier to foresee a lasting digital preservation and permanence. Furthermore there are some general concepts to ensure that digital files will be able to be used and kept for a longer amount of time. The first and perhaps most important is the idea of LOCKSS (Lots of Copies Keeps Stuff Safe) developed by the MetaArchive and is pretty well self-defined (Howard 2008, 16). Keeping as many different high quality formats and copies of documents on multiple hard drives and magnetic storage tapes is the best route to maintain the digital collection.

The fear of format/technological obsolescence is something to maintain in the back of your mind at all times, staying abreast to the newest technology trends and migrating files to new formats when they are developed is key to avoiding a total loss of digital files. Also, it is important to catalog the digital files with as much if not more metadata than a traditional analog object. Having the document retrievable by search is the best way to make sure it does not get lost. The time spent on digitization and also the software and components to create high-quality files is costly, it is important to see the digitization effort as a positive investment.

The digital movement is happening with great speed and it is important as librarians we stay on top of the movement so we maintain an active and new appearance amongst the public. As Ingrid Mason puts it in her article, "Virtual Preservation: How has Digital Culture Influenced Our Ideas about Permanence? [...]" Asserts:

> It is important to acknowledge these inherent tensions in any response to digital culture. The rate of change, the volume of digital material being published and the diversity of digital technology and digital culture overwhelm the possibility of applying the same level of human intervention as with analog practice. It is no longer possible to maintain the level of manual processing and to achieve the same comprehensiveness in collecting… (202).

It is understandable to be 'overwhelmed' by the surplus of digital content and also the public's outcry for more. Digitization efforts require a team of support and should not rest all on one person's shoulders; there needs to be cooperation between the librarians, technical services, and IT to make the entire process work. The important thing to keep in mind is to try and take the best steps forward in digitization and embrace this new medium.

Glossary of Terms from the SAA Glossary of Archival Terminology

By Richard Pearce-Moses

Born analog - Information that was created in a non-digital format and subsequently digitized.

Born digital - Information created in electronic format.

Electronic records - can encompass both analog and digital information formats, although the term principally connotes information stored in digital computer systems. 'Electronic records' most often refers to records created in electronic format (born digital) but is sometimes used to describe scans of records in other formats (reborn digital or born analog). Electronic records are often analogous to paper records; email to letters, word processing files to reports and other documents. Electronic records often have more complex forms, such as databases and geographic information systems.

GIF - Graphic Interchange Format. CompuServe developed GIF in the 1980s. Its palette is limited to 256 colors, but which 256 colors may vary from image to image. The image uses lossless compression to minimize file size.

JPEG - A standard (ISO/IEC 10918) that specifies a digital graphic file format that can reproduce a large color space and that can compress the data to minimize the file size.

PDF - The PDF format can be created using a variety of authoring tools, including Acrobat. The format can be read using the freely distributed Adobe Reader (previously called Acrobat Reader), which are both a stand-alone application and a plug-in for browsers. PDF is not perfectly platform-independent, as it requires a reader to render the file, although reader software exists for most platforms. PDF-A is an extension of the format intended to be appropriate for the long-term preservation of digital documents.

TIFF – Tagged Image File Format. A standard (ISO 12234-2) for storing a raster graphic and metadata that describes the image content and characteristics.

Bibliography

"ALA Definitions of Digital Preservation." Association for Library Connections & Technical Services, http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.cfm (accessed April 25, 2009).

Baldwin, Gil and George Barnum. "Government documents for the ages." American Libraries 32, no. 11 (December 2001): 38.

"BBC News – Technology – How the CD was developed" BBC News, http://news.bbc.co.uk/2/hi/technology/6950933.stm (accessed April 25, 2009).

Berry, John N. "Digital democracy? Not yet!." Library Journal (1976) 125, no. 1 (January 2000): 6.

Block, Debbie Galante. "The pluses and minuses of VR and DVD." EMedia (Medford, N.J.: 2002) 16, no. 9 (September 2003): 22-7.

Bradley, Kevin. "Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections: Strategies and Alternatives." Paris: UNESCO, (September 2006).

Collen, Lauren. "The Digital and Traditional Storytimes Research Project: Using Digitized Picture Books for Preschool Group Storytimes." Children & Libraries 4, no. 3 (Winter 2006): 8-18.

"Digital Preservation." Library of Congress, http://www.digitalpreservation.gov/ (accessed April 25, 2009).

"Digital Preservation Definition Announced." American Libraries 39, no. 7 (August 2008): 15.

Howard, Rachel. "Preservation Perspectives: Preserving Digital Information." Kentucky Libraries 72, no. 3 (Summer 2008): 16-17.

“Internet Archive: Wayback Machine.” Internet Archive, http://www.archive.org/web/web.php (accessed April 25, 2009).

Levi, Yaniv. "Digital Preservation: An Ever-Growing Challenge." Information Today 25, no. 8 (September 2008): 22.

Mason, Ingrid. "Virtual Preservation: How Has Digital Culture Influenced Our Ideas about Permanence? Changing Practice in a National Legal Deposit Library." Library Trends 56, no. 1 (Summer 2007): 198-215.

“Microsoft Open Specification Promise.” Microsoft, http://www.microsoft.com/interop/osp/default.mspx (accessed April 25, 2009).

Openoffice.org. “Open Office.org – The Free and Open Productivity Suite.” OpenOffice, http://www.openoffice.org/ (accessed April 25, 2009).

PBS Wired Science. “Video Luis von Ahn PBS.” PBS, http://www.pbs.org/kcet/wiredscience/video/284-luis_von_ahn.html (accessed April 25, 2009).

Pearce-Moses, Richard. “A Glossary of Archival and Records Technology.” The Society of American Archivists. http://www.archivists.org/glossary/ (accessed April 25, 2009).

Rog, Judith. “PDF Guidelines: Recommendations for the Creation of PDF Files for Long-term Preservation and Access, version 1.7.” The Hague, The Netherlands: National Library of the Netherlands (May 2007).

Turpening, Patricia K. "From Sheepskin Binding to Born Digital: One Hundred Years of Preservation in "Law Library Journal"." Law Library Journal 101, no. 1 (Winter 2009): 71-94.

Turro, Mireia Ribera. "Are PDF Documents Accessible?." Information Technology and Libraries 27, no. 3 (September 2008): 25-43.

Weyhrich, Steven. “Apple II History.” Apple II History, http://apple2history.org/ (accessed April 25, 2009).