Archival standards for born-digital documents:
Recommended methods for keeping stable preservation copies

Overview

As part of our plans to preserve student theses, dissertations, and newer editions of faculty texts and other culturally/academically significant documents, we inevitably will be tasked with preserving an increasing number of documents that originated electronically. These types of documents have been authored using various types of word processing and digital publishing software for decades, but the common practice had continued to be to print the final copy, and refer to the paper form as the final, finished product; the master original. Consequently, digital preservation would consist of scanning these analog objects back into a digital form, preserved electronically as scanned surrogates. Until very recently, we envisioned that scanning and digitizing from analog would comprise the bulk of how we digitally preserved all of our documents.

However, the increasing use of web-based publishing, online journals, and essentially paperless production has highlighted the benefits of seeking out the born-digital masters of preservation-worthy items whenever possible. Doing this affords us some advantages; namely, we can store the original in its most efficient digital form, often requiring less overhead and disk space while doing away with the quality challenges associated with scanning.

On the other hand, born digital preservation brings with it new challenges. Development of preservation standards for analog objects proved to be relatively simple, as the imaging industry laid much of the groundwork for us in terms of standardization across platforms. Further, development of future standards for digitized images, sound and video continues in an organized and orderly fashion, giving us plenty of time to contemplate migration to newer and better preservation formats.

Unfortunately, the same cannot be said for born digital documents. File formats for such objects vary widely, and the responsibility is upon us to identify a uniform set of file formats that we can adopt for preservation purposes.

As a result, a strategy for born digital document preservation must be adopted and followed that accomplishes the following:

- **Accurately renders** the formatting and content of the document, as intended by the creator of the document
- **Maintains stability** of the file format as well as possible. This may involve converting the document to archival formats, and storing both the original and the converted surrogate file.

Proposed Preservation Format Strategy: Multiple standards in play

Historically, born digital documents have been authored using a variety of different software packages, each with their own proprietary file formats. Early on, programs such as Wordstar, Wordperfect, Microsoft Works, ClarisWorks/AppleWorks, Adobe PageMaker, Quark Express, and others were distributed throughout the electronic document landscape.

More recently over the past decade, Microsoft Office has emerged as a de facto standard for general usage, with most businesses using it to create and distribute common document types. This usage has

resulted in a trickle-down effect to the consumer level on home computers and in academia as well. MS Office isn't perfect, however. The file formats used by Microsoft have evolved over the years as new versions have been released, and inconsistencies exist between versions in how document formatting is rendered.

At present, there are a number of formats developed by various consortia that attempt to solve the problem of maintaining a persistent document standard, and Microsoft itself has sought to modernize and make their document formats a formally accepted industry standard. Some of the more prevalent solutions include:

- **OpenXML:** A standard developed and endorsed by Microsoft and a consortium of other commercial software vendors, and is the standard document format used in the Microsoft Office suite beginning with Office 2007. These documents are often recognizable by their .docx, xlsx, and .pptx extensions.

- **OASIS OpenDocument (ODF):** An existing, open standard for file formats in use primarily in open source and "non-Microsoft" environments. These file formats are the default for OpenOffice.org and similar Free Software alternatives.

- **Portable Document Format/Archival (PDF and PDF/A):** A well-established standard with roots in Adobe PDF, a subset of which is now an ISO standard and a Library of Congress recognized format for digital document preservation.

There is also significant prevalence of legacy standards, a majority of which consists of legacy MS Office document types (.doc, .xls, .ppt, etc.) as well as more complex file formats for more intricate or specialized document types (LaTeX, Adobe InDesign, Illustrator, etc.). And finally, there are a multitude of document authoring platforms that are currently supported but have smaller market shares, such as Apple's iWork, current versions of Corel WordPerfect

Our choice of standards are based the ability to endure as technological advances continue to develop, and a widespread acceptance is key to ensuring easy migrating to newer standards when the time comes to retire existing choices.

The Recommendation: Our best case to preserve born digital documents while retaining longevity

Considering the state of the born digital document landscape as outline above, it is thus advisable that more than one preservation datastream for born-digital objects is utilized when possible. This strategy permits us to build redundancy into our repository, and ensure that regardless of whether one standard "wins out" over the other, our objects will remain with at least one relevant archival datastream. With that in mind, our strategy can be outlined as follows:

1. **Store the original document in its native format** when possible.
   In most cases, this will be an MS Office document, or a file from a similarly well-known software package. In some instances, the document we receive may already be rendered as a PDF file, in which case Step 2 below may not be necessary.

2. **Store an additional surrogate master in the form of a PDF/Archival file.**
   Most modern document authoring software, including MS Office and OpenOffice.org, have a

built-in capability to accurately "export" a document into a PDF version. This capability should be used when available to generate a faithful PDF file. Otherwise, the PDF/A can be generated using software available on RUcore platform.

Why PDF/A: An established standard to augment object datastreams

Although Portable Document Format has its roots in a proprietary system, recent efforts have proven fruitful – mainly thanks to Adobe, the creator of the file format – to have it recognized as an archival standard. PDF/A is defined by ISO 19005-1:2005, an ISO Standard that was published on October 1, 2005. According to the Library of Congress: "PDF/A is suggested as a preferred format for page-oriented textual (or primarily textual) documents when layout and visual characteristics are more significant than logical structure."[1]

The openness of this format has permitted a widening selection of software solutions to create archival PDFs from most digital documents. As indicated earlier, PDF "export" capability now exists on the market leading packages. Additionally, some computing platforms, namely OS X for Apple Mac computers and Linux environments, have a similar "print to PDF" feature standard as part of the operating system. Finally, free viewers exist for desktop and mobile computing platforms. This heavy documentation and wide accessibility make PDF/A a natural choice for acting as platform-independent method for preserving and making accessible born digital documents, without requiring users to purchase expensive, proprietary software to view the content.

Review provisions for special cases

The diversity that exists among born digital document formats virtually guarantees that a single standard will not address all use cases. In particular, this standard will not be well-suited to born digital documents that are formatted in such a way that a page-based presentation approach would be detrimental. In such a case, a review of how these documents were constructed will have to be undertaken, and the Digital Data Curator will need to consult the Cyber Infrastructure Working Group (CISC) and related subgroups on the best way to proceed.

[1] http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml