RUcore
Rutgers Community Repository

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

Introduction

This document describes recommended formats and best practices for the handling of born-digital documents. By "born-digital," we mean documents that did not originate in an analog or physical form (i.e. pen and paper, traditional typewriter or other physically printed format), but instead was created electronically, as a digital file. This standards document, for example, was created on a computer using Microsoft Word, and the "original version" of the document is a file residing on a hard disk drive. As a result, this document is born-digital.

Accepted File Formats for Preservation

- **Microsoft Office Open XML Format (MS-OOXML):** All Microsoft Office documents made with MS Office 2007 or later, and saved with 4-character file extensions, are OOXML documents.
  - Examples: .docm, .docx, .pptm, .pptx, .xlsm, .xlsx

  These documents are XML-based with an openly-published data structure.

- **Portable Document Format/Archival (PDF/A):** This is a subset of the common PDF format intended for archival preservation. Microsoft Office and most common PDF creation programs are capable of creating PDF/A files.

  It is strongly recommended to store the original source document (e.g. the MS Office file used to create the PDF) along with the PDF/A file whenever possible.

- **OpenDocument Format (ODF):** This an open source, royalty-free, XML-based document format commonly used by free alternatives to Microsoft Office, such as OpenOffice.org, LibreOffice, and Google Docs.
  - Examples: .odt, .fodt, .odp, .fodp, .ods, .fods, .odg, .fodg

- **Text files, and markup files formatted in text form (UTF-8)**
  - Examples: HTML, LaTeX, RTF, TXT, XML

Provisionally Supported Formats for Preservation

- **Binary (older generation) MS Office Documents:** Typically created by older (pre-2007) versions of Microsoft Office, and contain three-letter file extensions.
  - Examples: .doc, .xls, .ppt, .pub

  These files are supported by current (as of 2015) software, but may see reduced support in the not-too-distant future. When possible, these documents should be converted to the current OOXML-format to better ensure their longevity.

Presentation Formats

- **PDF and PDF/A**
- **Electronic Publication (ePub)** in cases where dynamic, device-specific formatting is permissible or desirable.

Background

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.

Recent technology changes have also resulted in an increased adoption of tablets, electronic book readers, and other low-power, light-computing mobile devices. Such devices are being used more frequently to access books and electronic documents, but have resulted in a spawning of new presentation formats. These flexible e-reader formats often redefine the page structure and layout of a document to suit the screen size and formatting of the specific device being used to access the content. The needs of those using these devices for access must also be considered when making digital content available online.

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 derivative 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.

3. **Provide a PDF/A presentation file, as well as an ePub file derviced from the master or user-generated, when the nature of the document permits it**

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 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.

ePub Format: Advantages and special considerations

ePub is an increasingly popular format for tablets, mobile devices and portable eBook readers. The advantage of this format is the ability to quickly reformat a text document in such a way that best fits the screen size and resolution of the device being used for access. ePub is built with the recognition that, unlike fixed sheets of paper, electronic readers come in multiple shapes, sizes and resolutions, and a fixed format for one size may not be appropriate for others. Additionally, some eReaders give the user

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

flexibility in how the ePub is presented. A variety of fonts, character sizes and color schemes are offered to permit the user to optimize the reading experience to their particular tastes.

This same set of features may, however, prove contradictory to some of the stated needs of certain collections in RUcore. Electronic Theses and Dissertations, for instance, have requirements set by academic departments which clearly specify the formatting and presentation of each document. In such cases, an ePub version of the same document may appear significantly different from these guidelines depending on the device used to display it, and may be deemed inappropriate as a presentation format.

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.