KEY ELEMENTS OF A FILE FORMAT STRATEGY

The only bad file format is one that hasn't been documented.

Tyler Thorsted

Brigham Young University
United States
thorsted@byu.edu
0000-0003-0292-0962

Within the Digital Preservation Community there are many references to policies on file formats, acceptable file formats, preservation policies and strategies, risk matrices, and action plans. All have the intention of defining and describing file formats and guiding decisions on which formats to preserve how, and when. My team and I originally created a File Format Action Plan, which was later migrated from OneNote to Confluence and then included more strategic plans for hundreds of file formats. This paper explores which key elements should be included in an effective file format strategy and the different ways such data can be used by people and systems. What works for one institution may not work for another, and the work created by a larger institution may benefit those with smaller resources.

Keywords – File Formats, Documentation, Registry

Conference Topics – We're All in this Together; From Theory to Practice.

I. INTRODUCTION

Recently I attended a webinar entitled, *"Do unacceptable file formats exist?"*.[1] The chat during the webinar was most telling in how everyone views the topic of file formats within their organizations. I observed that Institutional polices and available resources end up driving or limiting most of the work in creating strategies. My response to the webinar question is this: "the only unacceptable file format is one that hasn't been documented."

II. THE PROBLEM AT HAND

As digital preservation professionals we understand the work we do is more than a backup.

"A backup is a short-term data recovery solution following loss or corruption and is fundamentally different to an electronic preservation archive." [2]

"Digital preservation combines policies, strategies and actions that ensure access to digital content over time."[3]

Ensuring access to digital content over time is a monumental task. The last few decades have seen a number of changes in the way we interact with our computers and devices. This has led to an explosion of software releases and just as quickly, that same software becoming obsolete. Recent trends in software subscriptions models keep digital preservation professionals working tirelessly to ensure this access.

Preserving a set of born-digital files from a previous decade can be daunting as format identification tools may not always be able to identify the format. The file format may not be documented anywhere on the modern web. It may take a bit of sleuthing to find samples in order to understand which specific software created the files.

While some file formats were designed to be easily understood, there are many binary and container formats which end up requiring qualified guesses on their origin and signature.

In one instance, I was documenting a proprietary format and I felt I had gathered enough samples to identify the header and which bytes indicated version. When I reached out to the developer to confirm, their response was, "Please, do not use any hex editor and do not try to analyze the binary data file." This type of attitude makes preservation and access difficult for many many formats, increasing the risk in preserving.

In contrast, another format I researched was popular for a short time in the 1990's, often bundled with scanning software. It was a raster image format which faded off into obsolescence. Although the specifications were made public at the time, all links had rotted and were not available in the WayBack Machine. I was finally able to track down a developer and they were happy to share a copy of the specifications! [5]

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121127

Documenting old and new file formats reduces the risk of obsolescence, and if shared, reduces duplicated efforts.

III. Key Elements

Files stored in a repository all have unique attributes and history. The extension is not the only element dictating how these files are identified, migrated, or rendered. Below are some additional key elements that can be included in a file format strategy.

A. Identification

File formats should be identified using tools which look closer at a file beyond the extension. File Format Signatures can change over time. PRONOM PUID's are often used as the standard identifier, but there are many other tools which can be used.

B. History or brief description

Record a little background on the file format and its use at your institution. Include a current status of the software and its support by the developer.

C. Registries

There are many registries which you can refer to. Build on these for your institution specific needs.

D. Version information

Each version of software will create new versions of a file format. Knowing which versions of a file format are compatible with corresponding versions of software is important for proper rendering.

E. Specifications

If specifications for the file format exist, a reference to them should be included. If the specifications are unpublished or proprietary, details about research can be documented here.

F. Software to open/render

List which software can open and render the file format. Rendering matters. Not all software will open a file the same way. [4]

G. Software for migration

Software used for migration or normalization can be different than what is used to render. This element can also list software to avoid as it may cause unwanted changes. Include a decision tree for when a file is migrated.

H. Software to extract key properties

Detail which software can be used to extract key significant properties from the file format and their use.

I. Significant Properties (TechMD)

List which properties of the file format are important to extract? A TIFF may be an excellent raster image format to preserve, but if compressed with LZW, it may present a higher risk. List minimum set of required properties per institutional policy.

J. Risk

Risk assessments or preservation levels of support documents are useful tools for guiding strategy. [6]

K. Software to validate

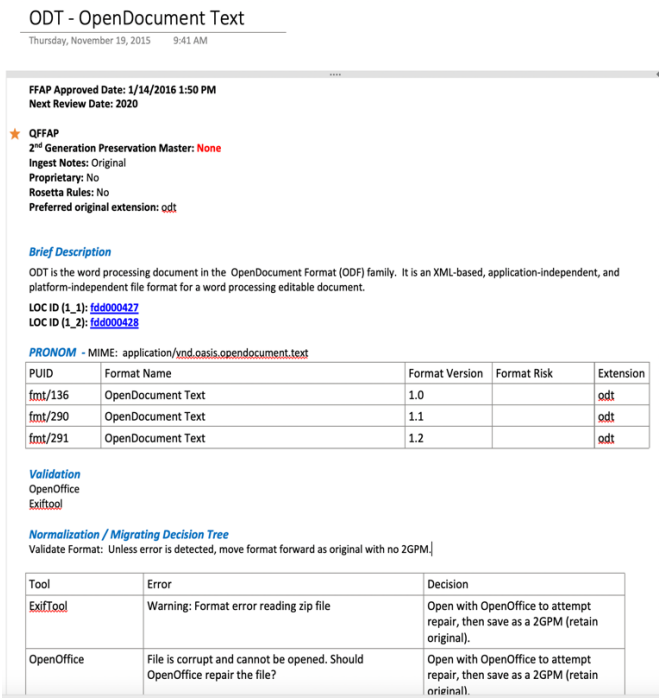
Many file formats can be validated to known specifications for institutional requirements. Software such as JHOVE or MediaConch can be listed here.

L. Rules

Many preservation systems have processing rules in place to help automate known identification and validation issues. Documenting these issues is important to understand decisions and preservation plans.

M. Platform (Mac/Win/Linux)

Some file formats and tools are platform-specific and require a certain environment to properly render or migrate.



ODT - OpenDocument Text

Thursday, November 19, 2015 9:41 AM

FFAP Approved Date: 1/14/2016 1:50 PM
Next Review Date: 2020

QFFAP
2nd Generation Preservation Master: None
Ingest Notes: Original
Proprietary: No
Rosetta Rules: No
Preferred original extension: odt

Brief Description

ODT is the word processing document in the OpenDocument Format (ODF) family. It is an XML-based, application-independent, and platform-independent file format for a word processing editable document.

LOC ID (1_1): fdd000427
LOC ID (1_2): fdd000428

PRONOM - MIME: application/vnd.oasis.opendocument.text

PUID	Format Name	Format Version	Format Risk	Extension
fmt/136	OpenDocument Text	1.0		odt
fmt/290	OpenDocument Text	1.1		odt
fmt/291	OpenDocument Text	1.2		odt

Validation
OpenOffice
Exiftool

Normalization / Migrating Decision Tree
Validate Format: Unless error is detected, move format forward as original with no 2GPM.

Tool	Error	Decision
ExifTool	Warning: Format error reading zip file	Open with OpenOffice to attempt repair, then save as a 2GPM (retain original).
OpenOffice	File is corrupt and cannot be opened. Should OpenOffice repair the file?	Open with OpenOffice to attempt repair, then save as a 2GPM (retain original).

Fig. #1, Example Strategy in Microsoft OneNote

IV. Audience

Who will be using this file format strategy? Is it just for preservation staff or is it intended for a broader audience? Institutional policies may be only useful internally, but documentation on file formats can be useful to share with the community.

V. Storing & Using the Data

Strategies can be documented in many ways. From simple Word Documents [7] to Excel spreadsheets [8], from Microsoft OneNote to Confluence. Others are using SQL databases or the popular Wikidata [9] ,Mediawiki approach. You can start small and grow the strategy over time or harvest from other sources into an actionable resource.

Digital Preservations Systems are moving toward more automated policies and preservation actions. These can be very useful, but don't let them replace your institutional strategies or be the only place such strategies are documented.

VI. Conclusion

Half the fun in documenting file formats is learning the history about the developer(s) and the purpose of each file format. Some were designed with the future in mind, while others were put together hastily to meet a deadline. Better still are the hidden meanings the developer left to be found by the curious (though, be careful of going down rabbit holes).

The statement, "The only bad file format is one that hasn't been documented" is not meant to convey that all documented file formats have no risk. It simply means that the more the community can document the formats in our repositories, the less risk they represent to preservation and access into the future.

1. REFERENCES

[1] A Panel Discussion: Do unacceptable file formats exist? February 9, 2023. http://bit.ly/3kVPNIn

[2] Digital Preservation: Continued access to authentic digital assets, JISC. http://bit.ly/3ystcGo

[3] "Definitions of Digital Preservation", American Library Association, January 18, 2010. http://www.ala.org/alcts/resources/preserv/2009def (Accessed March 8, 2023)

[4] Rendering Matters - Report on the results of research into digital object rendering, January 3, 2012 http://bit.ly/3kZUPUm

[5] XIFF File Format Research. http://bit.ly/3ZNZF5Q

[6] U-M Library's Digital Repository Services Registered Formats and Support Levels, http://bit.ly/3mAbqOY

[7] Strategies in Word Example. http://bit.ly/3F9fzA1

[8] NARA Risk Matrix in Excel. http://bit.ly/3JotggI

[9] Wikidata as a digital preservation knowledgebase, http://bit.ly/3mALWBd

[10] Just Solve the File Format Problem Wiki, http://bit.ly/3myBhqp