White paper

Digital Preservation White Paper Series

Digital Preservation Policy Creation

Preservica

1. Introduction

The scale and diversity of digital information to be preserved necessitates the development of digital preservation policies and strategies to ensure the long-term accessibility, authenticity, and usability of digital assets.

While a complete digital preservation policy will encompass many facets, potentially including collection strategies, documentation and metadata standards, storage choices and replication policies, and legal and ethical considerations such as copyright and licensing, this whitepaper focuses on the expression of digital preservation policies that can be defined and executed within digital preservation systems, in particular file format preservation policy.

2. What is in a format policy?

As explored in the whitepaper 'Automated File Format Preservation,' the key starting points for a preservation policy are knowing what you have, and knowing how you want to use it. File format identification and characterization are critical to the former, while knowing how you want to use your data requires an understanding of the wants and needs of your current and future designated communities.

A format policy will specifically consider:

- The file formats within your collection, plus those that are anticipated to be deposited to your collection in the future
- The technical characteristics or properties of the files, potentially considering specific codecs, color profiles, character encoding, password protection or encryption, and any other technical properties that may warrant further consideration beyond the base file formats
- The physical make-up of those files, such as whether they are standalone file formats, multipart assets where relationships to other files are critical to retain, or archival containers that may need to be unpacked and have their own contents individually assessed
- File format migration strategies
- File format rendering capabilities
- Sensitivity and access control. While critical regardless, when considering file format policy this may involve the creation of differing migration and access strategies for different audiences.

Preservica provides hundreds of business rules that allow for file format migration, rendering, validation, and property extraction. Preservica's team of digital preservation experts continually research the digital preservation landscape and add new business rules with every new Preservica release. Business rules enable the expression of file format policy through the creation of rulesets that can be applied across your digital collections.

3. Sources of information

3.1 PRONOM

PRONOM [1] is a registry of data about file formats, published and maintained by The National Archives in the United Kingdom. PRONOM contains data on over 2,300 file formats and continues to grow through community submissions. PRONOM's key feature is the provision of

‘file format signatures’, machine-interpretable descriptions of elements within file formats that can be used to reliably identify the exact file format for a given file instance, and to assign a PRONOM Unique Identifier (PUID) to each file to describe its format. While PRONOM avoids offering opinion on file format policy, the PUID scheme has become a common basis for digital preservation systems to apply preservation policies at the file format level, enabling the association of file formats, with actions (such as file format migration, validation, or rendering), and the tools (such as code, software libraries, or third party applications) to execute those actions.

3.2 Community Exchange

The digital preservation community is full of friendly, experienced practitioners who are willing to share ideas, experiences, and solutions. There are many mailing lists, social media communities, and community web forums open for allcomers to join. There are several digital preservation-focused conferences throughout the year and across the globe, where theory and practice are shared and discussed. There are also often more local or industry-specific special interest groups, with themes such as ‘web archiving’ or ‘business archives.’ Even where formal groups don’t yet exist, it can be useful to reach out to peers in similar institutions to your own to start to build those networks of interest and to share digital preservation challenges and file format policies.

3.3 Library of Congress

The Library of Congress publishes their own file format registry. While PRONOM focuses primarily on the challenge of file format identification, the Library of Congress Format Description Registry [2] provides extremely thorough descriptions of over 500 file formats. The registry contains detailed descriptions, including detail of file format risk factors, such as sustainability, ubiquity, licensing and patents, and other dependencies, so can serve as a useful starting point for understanding format risk within the context of your own institutional risk appetite, that may in turn influence your file format policies.

3.4 National Archives and Records Administration (NARA)

NARA publishes a range of file format related resources. For instance, as part of their Transfer Guidance for federal agencies in the US, they publish a list of file formats that are recommended or acceptable for transfer to NARA [3]. NARA also publishes Preservation Plans for hundreds of file formats, which they make available as Linked Open Data [4]. These detailed plans cover whether a given format should be retained as-is, or transformed (migrated) to a different format, and where appropriate specifies the preferred output format for migration. These resources will likely influence the format policies of federal agencies that need to transfer data to NARA but can be referenced by any organization.

3.5 WikiData

WikiData [5] is a collaboratively edited community knowledge base for structured data that serves data for over 100 million items. A significant effort to utilize WikiData as a digital preservation knowledgebase was spearheaded by Yale University Library and has continued to be shaped by a legion of committed volunteers. A dashboard and editorial frontend, WikiData for Digital Preservation [6] was also established. The aims of the effort include describing the relationships between file formats and the tools that support them. This in turn has helped to support Emulation as a Service efforts to simplify the process for using emulation as a digital preservation strategy.

3.6 Preservation Action Registries (PAR)

PAR [7] is a project that aims to support the sharing of technical information on digital

preservation across the global digital preservation community and between digital preservation systems.

PAR is a collaborative project established by JISC and includes Preservica, Artefactual Systems, and the Open Preservation Foundation.

PAR provides a JSON schema and an API specification that facilitates the sharing of digital preservation policy, with an aim to establish a set of trusted sources that can provide the catalyst for information and good practice sharing across a set of proprietary Digital Preservation systems as well as the wider community.

3.7 International Comparison of Recommended File Formats (ICRFF)

The ICRFF dataset [8], published by the Open Preservation Foundation (OPF) is a resource that collates publicly available file format policies into a downloadable spreadsheet. It is not intended to be used as a policy document, but it can serve as a useful insight into the file format preservation policies employed by a diverse range of institutions.

Preservica's digital preservation experts are regular contributors to many community initiatives and continually monitor community resources for opportunities to enhance the Preservica system. Preservica also provides a dedicated community forum to enable our users to discuss digital preservation challenges with like-minded peers.

4. Policy elements

4.1 Intended future usage

As described in the Preservica whitepaper 'Automated File Format Preservation', the needs and requirements of your designated communities will influence your file format preservation policy. Are they digitally confident enough to work with data in its original format? Will they expect a modern-day facsimile? Will they just want to view the document through a web browser? Are there different audiences with differing needs and expectations? The answers to these questions will inform a file format policy and may result in the need to create multi-faceted policies covering many use cases.

4.2 Does a "future proof" format exist?

File format obsolescence is hard to predict. Formats with interesting technological innovations have come and gone, while some relatively technologically basic formats have stood the test of time and continue to see heavy use today.

In the 1990s, during the early days of the worldwide web, the Graphics Interchange Format (GIF) was a well-established image format, but Unisys, then-holders of the patent for the LZW compression algorithm at the heart of GIF wanted to monetize their patent [9]. They proposed to charge royalty fees to software developers utilizing LZW technologies. This was unpopular, and many software creators actively discouraged the use of GIF. There was even a 'Burn All GIFs' boycott movement. Meanwhile the Joint Photographic Experts Group (JPEG) developed a patent-free alternative to GIF that would remain wholly free to use, the Portable Network Graphics (PNG) format. The patent in question expired in 2003, and despite GIF's limitations such as its limited color palette and lack of support for transparency, GIF's use as a lightweight animated graphics format became a cornerstone of the web throughout the early part of the

21st Century. Today, nearly 40 years after its creation, GIF remains a hugely popular format on the web.

Since its original standardization in 1992, the JPEG image format has seen many file formats challenge its primacy as the photographic image format of choice for the web at large [10]. Many image formats have emerged promising more efficient processing, more efficient or lossless compression, and better image quality. Even JPEG's creation and maintenance committee, the Joint Photographic Experts Group, has published specifications for formats intended to either complement or succeed JPEG, including JPEG LS, JPEG 2000, JPEG XT and JPEG XL. Alternatives such as Better Portable Graphics, WebP, and HEIF have also emerged. For now, JPEG remains one of, if not the most popular image format on the web. JPEG also had its own patent-related controversy in the early 2000s.

Developed in the mid-90s, the Fractal Image Format used a clever form of image compression based on fractals that facilitated upscaling, a process to increase the resolution of an image. This process worked very well on photographs of the natural world, such as landscapes and textures, with the compression algorithm able to reasonably predict and interpolate additional image data while retaining sharpness. The fractal image technology was licensed to Microsoft for its Encarta encyclopedia application, and a plugin was also created for Adobe Photoshop, but the core Fractal Imager software was discontinued in 1998 [11] and today it is difficult to find software that will work with the image format. Today, the use of AI rather than fractal technology to predict and interpolate image data for upscaling is becoming commonplace, but it shows that the fractal image technology was a potentially viable solution to a common problem if only it had been more well known at the time.

So, we see that innovative uses of technology do not ensure a given file format will be 'future-proof' and we also find that technical limitations, legal challenges, and even attempts to deliberately replace an older format with a modern alternative will not necessarily succeed. This stresses the importance of finding and retaining solutions for identifying, migrating, and rendering file formats while they are in contemporary use, as the technological landscape can shift considerably in an extremely short space of time.

4.3 Preservation vs Access

Different file formats suit different purposes, so it is common to have different format policies for preservation and for access. For example, the Material Exchange Format (MXF) is typically used for broadcast-quality video data, however files tend to be extremely large therefore are usually unsuitable for delivery over the web to consumer-level end-users. MP4, using the H.264 codec and optimized for web delivery allows for quick start streaming, with much smaller overall file sizes, however this compression is typically lossy and is much more suitable for access rather than as a preservation representation.

It therefore follows that it is perfectly acceptable to have a format policy for access-related purposes, and a policy for preservation-related purposes. A digital preservation system should allow both to co-exist and produce multiple representations to meet these aims.

4.4 Migration vs Render

JavaScript libraries are available for many file formats that allow for them to be rendered natively within a web browser without the need for a preliminary migration, or in some cases it may be preferred to transcode data to a renderable format at the time of request rather than up-front, should the transcoding process be performant and efficient enough. As such the choice of available rendering options may influence a file format policy as it may be preferred to rely on such rendering tools.

4.5 Available Tools

Preservation actions such as migration, validation, rendering, or property extraction, rely on software code to perform the intended actions. This code may be provided by third-party tools, native libraries, or bespoke software created in-house, but it is certain that over time the software will evolve. Security issues will be patched, new Application Programming Interfaces (APIs) will be introduced, and support for certain formats will improve or will sometimes even decline or be withdrawn altogether, for example the removal of support for Adobe Flash in most web browsers. New tools will emerge that provide better support for existing or new formats and allow for a greater pool of preservation actions.

A digital preservation system therefore needs to be able to adapt to this changing tool landscape. When new tools are introduced, it must be possible to associate preservation actions with their intended file formats. In cases where a new tool supersedes an older tool and provides some further functionality, such as a better format migration output, or a wider range of properties to be extracted, format policy should be executable to take advantage of the new functionality.

4.6 Setting up the rules

Preservation rules will typically include the following elements:

- Tools – the software tooling that will perform an action
- File formats – the file formats that preservation rules will be performed against
- Preservation actions – the specific parameters that a tool will execute to perform the intended action. Preservation actions may also be grouped by the type of action intended, for example 'property extraction' or 'migration' actions
- Business rules – the association between file formats and the preservation actions to be performed. Business rules may differ according to an organizational preference, for example one institution may prefer to use Jhove for property extraction for a given format while another institution may prefer to use ExifTool
- Rulesets – the collection of business rules that are to be applied to a given context. Rulesets may, individually or as a collective, form a format policy.

For example, a tool (e.g. ImageMagick's 'Convert' utility), can be used with a file format (e.g. fmt/353 - Tagged Image File Format), by performing a Preservation Action (e.g. 'migrate-to-jpeg', expressing the command 'convert <input.tif> <output.jpg>'). The Preservation Action is associated with the Business Rule (e.g. 'jpeg-migration-imagemagick') which defines a list of file formats that can be migrated using the Preservation Action. A Business Rule can be executed manually, or automatically by policy using a 'Ruleset' (e.g. 'default-image-migration').

For the most part, where appropriate tools are already available and the file formats are known to the digital preservation system, it should be well-defined and relatively straight-forward to create new rules and to add these rules to a ruleset, then to apply rulesets as file format policy items.

4.7 When do rules apply?

Typically, policy rules will be initially evaluated and applied either at the point of ingest to a digital preservation system, or as an asynchronous task to happen shortly after ingest. It should be possible to apply a new policy rule at any point post-ingest, and for the new policy to be applied where appropriate across the relevant digital collections.

4.8 Where do rules apply?

While some format policies may be appropriate to apply to all files within your digital collection, there may be instances where you want to limit your policies to certain subsets of collections. For instance, you may wish to have a different output migration format for different sets of records depending on their location within a folder hierarchy or depending on their intended audience and managed through security access controls.

4.9 Handling change

Change is inevitable. The format policy you create today will necessarily evolve over time. Perhaps a brilliant new file format is released that your end users insist upon for content delivery. Perhaps a hitherto popular and well-supported format is found to have an alarming security vulnerability and is no longer considered safe to use. Perhaps a regulatory body for your sector decrees that all files must be published in the Open Document Format without exception.

Any system that acts on format policy must be able to react to change to that policy. This should mean that a user with appropriate permissions is able to create a new policy whether to enhance or replace an existing policy, that the system is able to report on the impact of a potential new policy, and that the system is able to apply a new policy en-masse with minimal user intervention.

Preservica provides a comprehensive and growing set of preservation actions, business rules, and rulesets out of the box that can be used to create file format policies easily and intuitively. Through the Preservation Actions Registry API it is easy to create more rulesets that suit your specific needs.

5. Example policies

5.1 Parsimonious Preservation

Parsimonious Preservation [12] is a policy suggested by Tim Gollins during his tenure as Head of Digital Preservation at The National Archives in the United Kingdom. It is a minimal intervention policy that argues that, although technological obsolescence is real, a more imminent threat is poor capture and failure to ensure safe custody of original digital materials.

The paper acknowledges the limited resources of many heritage institutions and argues in favor of focusing on data capture and data storage over attempting to predict or pre-empt technological obsolescence, to maximize the return on preservation effort where resource is limited. The paper recognizes that this policy will primarily be of use for those institutions preserving the normal administrative data of an organization, or smaller institutions starting their digital preservation journey in what can be an overwhelming field, but that this will not suit institutions with large archival holdings or specialist data.

5.2 Migrate to Open Standards, such as Open Document Formats

Some format policies focus on the perceived improved accessibility provided through the use of Open Standards, the idea being that by making data available in file formats that follow open standards, barriers to access, particularly those associated with the costs of proprietary tools, are broken down, since end users will usually be able to make use of free and open-source software tools to access published data.

The Cabinet Office department in the United Kingdom published a policy paper for use across the UK Government on Open Standards principles [13] that emphasizes the benefits of open standards including improved interoperability between IT systems and reducing the likelihood of storing duplicate data through standardization. This policy will therefore likely influence the nature of data that is produced today across the UK Government, but may also have an impact on legacy data that may have been previously produced using proprietary licensed software, since it follows that files may need to be migrated to Open Document Formats in order to meet the aims of the policy when it comes to publishing data for end-user consumption.

5.3 Migrate to the business formats used by your institution

The Microsoft 365 ecosystem has become dominant across many different industries. Office 365 reportedly had over 380 million 'commercial seats' as of April 2023 [14], and upwards of 60% of Fortune 500 companies reported as using Microsoft 365 [15]. It is therefore likely that most of these companies will today routinely work with Microsoft Office formats and that migration to Microsoft Office formats may be a sensible choice.

The main competitor to Microsoft for online productivity software is currently Google, with its Google Workspace offering. The native Google Workspace file formats are not intended for document exchange outside of the Google Workspace ecosystem, however both Open Document Formats and many Microsoft Office formats can be worked with by Google Workspace so either may be suitable format families for migration output.

Many legacy word processor or spreadsheet file formats may not be able to be natively opened with either of these systems, therefore a format policy that migrates to a suitably compatible set of formats will be of benefit.

5.4 Migrate to popular, easily accessed formats

Formats such as PDF (for document and document-like formats), MP3 (for audio), MP4 (for video) and JPEG (for images) are extremely common across the web and there are many options for viewing these file formats, including within the context of a web browser. As such these may be a useful choice for migration output, particularly where access is a primary concern.

Preservica supports hundreds of options to enable you to build a format policy that works for your institutional needs. Preservica's patent pending Automated Digital Preservation technology facilitates the application of file format policy at scale since it allows you to express your policy and allow the system to take care of ensuring your digital assets conform to your policy.

6. Further Reading

The Preservica White Paper Library covers all aspects of Digital Preservation showing the strategies and solutions required to ensure information is available in the future.

The following are also sources of information on the topics discussed:
[1] PRONOM, https://www.nationalarchives.gov.uk/PRONOM/
[2] LIBRARY OF CONGRESS FORMATS DESCRIPTIONS, https://www.loc.gov/preservation/digital/formats/fdd/

[3] NARA TRANSFER GUIDANCE, https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html
[4] NARA LINKED DATA, https://www.archives.gov/preservation/digital-preservation/linked-data
[5] WIKIDATA, https://www.wikidata.org
[6] WIKIDATA FOR DIGITAL PRESERVATION, https://wikidp.org
[7] PRESERVATION ACTION REGISTRIES, https://parcore.org
[8] ICRFF, https://openpreservation.org/news/new-community-resource-international-comparison-of-recommended-file-formats/
[9] BBC, 20 JUNE 2003, GRAPHICS FORMAT WINS FREEDOM - http://news.bbc.co.uk/1/hi/technology/3007862.stm
[10] WIKIPEDIA, JPEG SUCCESSORS, https://en.wikipedia.org/wiki/JPEG#Successors
[11] THE INTERNET ARCHIVE, ITERATED SYSTEMS FRACTAL IMAGER PRODUCT PAGE, 26 JUNE 1998, https://web.archive.org/web/19980626190724/http://www.iterated.com/products/fractalviewer.htm
[12] PARSIMONIOUS PRESERVATION, https://cdn.nationalarchives.gov.uk/documents/parsimonious-preservation.pdf
[13] GOVERNMENT DIGITAL SERVICES BLOG, REFRESHING THE OPEN STANDARDS PRINCIPLES, https://gds.blog.gov.uk/2018/04/09/refreshing-the-open-standards-principles/
[14] MICROSOFT TECH COMMUNITY FORUM POST, MICROSOFT FY23 Q3 RESULTS HIGHLIGHT CLOUD SUCCESS, https://techcommunity.microsoft.com/t5/microsoft-365/microsoft-fy23-q3-results-highlight-cloud-success/m-p/3806551
[15] MICROSOFT BLOG, MICROSOFT SOLUTIONS BOOST FORTUNE 500 FRONTLINE PRODUCTIVITY WITH NEXT-GENERATION AI, https://blogs.microsoft.com/blog/2023/08/09/microsoft-solutions-boost-fortune-500-frontline-productivity-with-next-generation-ai/

7. Other papers in the Preservica expert series

Digital Preservation Overview

Automated File Format Preservation

Preserving Multi-Part Information Assets

Digital Preservation Metadata