White paper

Digital Preservation White Paper Series

Automated File Format Preservation

Preservica

1. Introduction

The vast majority of information created today is digital, and a significant proportion of this needs to be retained for regulatory compliance, legal protection or value creation. However, unlike physical information on paper or film, digital information is saved in a form that is not directly usable by humans and requires software running on computers to allow humans to access it. The structure of the information has to conform to the expectations of the program reading it in a way that tends to be much stricter than the leeway permitted when humans look at a physical object – we can cope with slight variations in layout in a way a computer program is unlikely to support.

When preserving information for the long term, this dependence between the raw information stored on a computer and the programs needed to interpret it is critical. However these programs often only run on a specific version of an operating system which itself runs on a specific hardware platform such as a server, laptop, or phone. When retrieving a file saved decades ago, it is easy to find that the program that reads it no longer runs on the computer devices of today. Even if the program still runs there are more challenges, for example lost license keys or access passwords or the need for specialist skills to use the software. Even if it is possible to access it using huge amounts of digital forensics effort, the expectations of instant access to information mean that the consumer may not be able to get to the information within the time span required.

All of this may mean the information is as good as lost. Automated File Format Preservation puts in place a range of strategies that reduce or eliminate these risks so the information can be accessed quickly and effectively in a form that fully represents the information to the quality required.

Preservica has been a pioneer in file format preservation for over two decades and has participated in many of the groundbreaking developments on which all digital preservation systems now depend. Preservica now offers fully automated file format preservation implemented by a team of experts who work with the wider community to ensure digital content is always in step with acknowledged best practice.

2. File Format Preservation Overview

To ensure the information is readable when it is needed digital preservation systems go through a series of steps to prepare the information and then deliver a number of strategies to deliver future readability. For any particular piece of information, you need to:

- Find out what you have. You need to know the format that the information is held in (both the format family and the version) and to get some of its properties so you know exactly what you need to do to care for it.
- Plan for how you might want to access the information in the future. How you preserve the information depends on the anticipated usage, for example is it read only or do you need to edit and re-purpose the content.
- Plan for how to preserve the information so that you can fulfill the use cases. You need to decide whether to migrate the information to another more modern format appropriate to its anticipated use, or to maintain software that allows users to access the copy you have got.
- Put it into practice. This will allow manual or preferably automated actions that run the processes required to implement to policy and do this on the large scales that are now becoming common in most preservation systems.

Later we will explore these steps in more detail and show how they can be automated at scale.

3. What is a File Format?

In computer systems information is stored on a disk as a series of zeros and ones, known as bits. Computer programs read these bits and make sense of them by using a set of rules that convert the zeros and ones into something the program can use for the process for which it was written. These rules define a format.

One example often used to illustrate this is a text file. One very common approach, particularly for speakers of languages using a latin alphabet, uses a set of rules called ASCII (American Standard Code for Information Interchange) to map certain bit patterns onto common letters, numbers and punctuation characters. Simple text editors could read and edit files because the meaning of the ASCII codes for each character are defined.

Even something as simple as this didn't stand still. ASCII started as a set of 7 bits, allowing up to 128 characters to be encoded, and was then extended to 8 bits, supporting up to 256 characters. Now it has been largely replaced by Unicode, which can encode up to 1,112,064 unique characters. 149,813 characters from all known languages, including emoji, have already be defined within this standard. This is most commonly encoded as UTF-8 (Unicode Transformation Format – 8bit), which uses between 8 and 32 bits to encode the entire Unicode character set, but other encodings such as UTF-16 and UTF-32 also exist. So now even a simple text editor first needs to know which type of character set the file is using, ASCII or one of the UTF sets.

Of course other types of data need more complex rules. Images need complex rules to map bit values to colors that can be presented on screen, and often include complex compression algorithms to reduce storage demands. Videos have two levels of format, one to say how the image and audio streams are arranged within the file, and another layer to say how the bits in a stream are encoded and how they can be transformed back into moving pictures and sound.

Some formats use a generic encoding to comply with specific rules. For example Extensible Markup Language (XML) is a generic format using UTF-8 encoding but is very often used to hold specific data, for example GPS Exchange Format (GPX). The rules for the specific use of XML are held in a separate schema definition but these can still be difficult to understand.

Some formats contain other files collected and often compressed. Common examples include ZIP and TAR. These can often be thought of as transient artifacts of the movement of files and folders between systems and are often unpacked when loaded into a preservation system. Some are unpacked but also include additional information which must be extracted, including ISO Disk Images and email MBOX files.

Some programs need multiple parts of information to describe the information they require, and this information is compressed into a single file. The most common example is the 2013 Microsoft Office suite, all of which hold multiple files in an XML format within a ZIP file. To test this out rename an office file to myfile.zip and have a look inside.

Some programs take a very strict approach to interpreting formats, and the data will fail to display if it does not conform to the specified format. Other programs expect errors and handle them gracefully, most commonly web browsers, which read Hypertext Markup Language (HTML) which very often does not conform to the standard.

The set of rules that define a file format can be agreed, published and maintained by a committee spanning multiple organizations, for example JPEG published by the Joint Photographic Experts Group. Some are defined by a single organization but still published so other organizations can use it, for example the current Microsoft Office formats. Some are proprietary, confidential and can only be interpreted by reverse engineering the bit patterns within a selection of files.

There are several registries of file format information covering a range of widely used formats. One of the most authoritative is PRONOM, created and maintained by the UK National Archives [1]. This is updated regularly and is used by most digital preservation systems as the reference for the formats they handle.

So, whilst the concept of a format is easy, the formats themselves quickly become complex and opaque. This makes preserving the information they contain complex, demanding specialist research and large-scale processing to ensure that the information lives on even if the format does not.

Preservica's team of Digital Preservation Experts understand the formats used for digital information at a profound level. They have contributed to research in this area and lead the way in understanding how this affects the required preservation strategies at an intellectual and practical level. Preservica experts developed the original version of PRONOM and contribute regularly to its maintenance.

4. Preparing digital assets for Preservation

4.1 Preparing an "Information Asset"

Before preserving some information, you must first understand what files are needed to represent it within the software. This "atomic" set of information is the smallest component that fully describes the information and in most cases this is a single file, for example a document, image, spreadsheet or video.

In an increasing number of cases more than one file is needed to describe a piece of information. Examples include an email with attachments, a tweet with embedded images or videos, and a SharePoint list item with data in columns and attachments in data files. These situations are described in detail in our "Preserving Multi-Part Information Assets" white paper in this series.

The information asset may also have metadata attached to give context and meaning to the asset. This is important for understanding the data and can also be used for search (finding aids). This is described fully in our white paper "Digital Preservation Metadata".

The rest of this paper is concerned with the preservation of information held in a single stream of bits, most usually held in a digital file, and all the steps required to ensure it is kept accessible in the long term.

4.2 Format Identification

To preserve the information in a file you need to know which format it is in. Nearly all computer systems rely on the suffix of the file name to decide on the format, so for example they assume "myfile.doc" is a Microsoft Word 97-2003 file. However, this is highly flawed. Firstly .doc is also used by many other file formats, not least all the versions of Word that preceded 97-2003 – the things you can do with Word 97-2003 are totally different to how you would handle Word for Windows 1.0 which also uses the .doc suffix. Secondly, anyone can change the suffix of a file – the system might at best warn you but it doesn't stop you.

The better way to identify the exact format of the file is to look inside the file at the binary and look for patterns that are unique to that format. A good example is PDF, which always has a set of characters in a fixed part of the file that spell out "%PDF-x.y" where x.y is the specific version number e.g. "%PDF-1.4". By searching for that pattern in that location we know this is a PDF

version 1.4 file. Other formats are more complex and have patterns throughout the file but the principle is the same.

As described above, some generic formats are used as the basis for more complex information, so for example parsing a GPX file would show it conforms to GPX and also XML. In this case the priority would be the most specific format, so we would assume it is a GPX file.

Also, some formats are made up of a set of files held inside a compressed file, so for example running a simple pattern match to identify a Microsoft Word 2013 file will tell you it is a ZIP file. Opening the ZIP file, you will find a specific set of contents that tell you it is actually a Word file as well as a ZIP file, but that as Word is more specific this is the one you should use. Identification of those file formats will use a two-stage parse, first to find the ZIP file match, and then to find files with other specific formats within that.

The PRONOM database run by the UK National Archives includes a tool called DROID which runs the pattern matches and compressed file examinations described to identify a file's specific format and version. The data and tool are freely downloadable are updated regularly and currently contain the format patterns for over 2000 file formats.

The Preservica Digital Preservation Experts developed the initial versions of DROID and PRONOM, and have contributed significantly to the maintenance of PRONOM's content, helping it grow to be the leading file format identification resource. Preservica has DROID built in to its ingest process so all files ingested into the system are run through DROID to discover their format.

4.3 Property Extraction

Knowing the format of a file is a useful first step to making sure it can be used long into the future, but to be really confident you need more information. Most files have measurable or readable properties that are useful when devising a plan for long term care, so extracting these properties can help ensure accurate and reliable future access.

The properties extracted depend on the type of file. Some comply to de-facto standards such as Exchangeable Image File Format (EXIF) which describes images, sounds and some other files. Others are proprietary but tend to be the same across multiple vendors, for example the number of pages and words in a document. Some are important in interpreting the file, for example the algorithm (codec) used to encode images and sounds in a video file. Others are useful context about the file, for example the author and copyright.

These properties are useful when deciding how to process a file. They are also useful when migrating from one format to another as some key properties should not change during the migration process, for example the number of pages in a document.

The digital preservation community has developed and maintained a number of open-source tools that can be used to extract file properties. These include ExifTool (Images, Microsoft Office OLE, ZIP, Rich Text), JHOVE (Images, PDF, HTML), veraPDF (PDF/A), Apache POI (Office Open XML), MediaInfo (Videos), and FiWalk (Disk Images) – there are more listed on the DigiPres Commons website [2]

In addition, preservation systems can develop their own tools to extract properties for more complex objects such as 3D as discussed in the White Paper "Preserving Multi-Part Information Assets".

Preservica wraps all of the property extraction tools mentioned covering documents, spreadsheets, presentations, images, videos, audio, emails and disk images. It also includes its own tools to extract properties from complex objects such as 3D models, emails with attachments and tweets. These are run every time a file is ingested into the system.

4.4 Format Validation

A file format is a strict set of rules that define how information is structured in computer terms. Given that, it should be possible to validate that any particular file conforms to the rules for the format that it is supposed to be written in, and this would be useful when ensuring that the file is preservable for future use. For many formats there are good quality file validation tools that can be run and will let you know if the file complies with the rules.

However in reality it can be surprisingly hard to get a definitive answer due to the complexity of the rules and the ability of programs to compensate for errors. With web pages encoded in HTML for example the number of pages that fully comply with the standard is fairly low but browsers can cope with the many deviations and still display the page in the way the author intended. Given that each browser will have different limits of which differences they can compensate for, validating HTML is pointless.

With ubiquitous formats like the Portable Document Format (PDF) family, the rules are so complex and the variations so infinite that even teams of experts struggle to say which variations can be ignored and which cannot. They also depend on the intended use, so for example in routine use it does not matter that the font definitions used in PDF are stored with the program and not with the file. However, when saving a file for decades it is better to embed the font definitions inside the file in case they disappear from regular use at a later date. Validating a PDF for long term retention thus has a different set of criteria for validating for immediate consumption.

Many of the tools used to extract properties for files also perform validation, most notably JHOVE (Images, PDF, HTML) and Apache POI (MS Office). The veraPDF tool used for PDF/A can use different profiles to say which rules it should warn about and which it should ignore. Some generic formats such as XML can be validated using tools built in to standard programming languages, and complex data objects can be validated using the tools built into advanced digital preservation platforms,

In the end Digital preservation systems have to take a pragmatic approach to validation, running checks for certain formats, but knowing when to reject content and when to warn that the files have an increased level of risk due to deviations from the ideal.

Preservica wraps all common validation tools as well as its own advance tools for complex multi-part objects. Non-conformance warnings are presented to the user to make the decision of whether to accept deviations from the standard.

4.3 Anticipating Change

Digital preservation systems identify and validate and extract properties for all the files that arrive on their system. However, IT is a highly dynamic domain and all things in it are subject to change. This can result in incorrect information being extracted from the content leading to

inappropriate preservation actions.

One obvious change is the adoption of new file formats. Recent examples of this are HEIF (High Efficiency Image File Format) used for photos and HEVC (High Efficiency Video Coding), also known as H.265, for videos. These were released as the default for images and videos on Apple devices from iOS 11 but the existing format identification tools were not ready and marked them as "unidentified". A new format and the pattern required to identify it was quickly released but in the meantime digital preservation datasets contained files which could not be used.

Another digital photography example is the Canon Raw series of formats used to hold unprocessed images from Canon cameras. The first version, CR1, was a proprietary format that was easily identified. The second version, CR2, is in fact encoded to the TIFF image standard so initially files were identified as this format. When systems tried to process them as TIFF they failed however as they were a specialist implementation of TIFF. The next version, CR3, conformed to the MP4 video standard allowing it to contain images or video, but of course could not be processed as standard MP4. Although new versions of the property database were released, the initial ingests of CR2 and later CR3 were incorrectly identified and could not be properly used.

Sometimes there are errors in the pattern rules in the format database. An example of this was the Word 97 Template format (DOT), which for a while was incorrectly identified as a Word 97 document (DOC). As above, an update to the format database was released, but in the meantime any Word 97 DOC file may actually be a Word 97 DOT file meaning it may not be presented correctly to the user.

These examples show that change is inevitable. Any digital preservation must be able to react to new formats and incorrectly identified files and retroactively re-process files to correctly prepare them for preservation.

Preservica's patent pending Automated Digital Preservation technology contains the ability to efficiently re-process formats that are now considered suspect and to re-apply the format preservation policy for files that are found to have the wrong format. The list of "at-risk" formats can be identified by the user's own experts or by Preservica's Digital Preservation Experts who can publish lists of recommended re-processing instructions that the system will automatically apply. By selecting the latter as the default, the user's content is always conformant to the current recommended best practice.

5. Planning for future usage

Having ingested and prepared the files you wish to preserve for the long term, the next task is to try to anticipate who the future users of your information are, and how you expect them to want to use it. These users are sometimes called the "designated community".

This is of course quite difficult as the future may be decades away, and there might be multiple different designated communities for the same information, but the earlier you are able to understand what use the material may be put to, the earlier you can get it in a form that supports that use, and the lower the risk you can't re-process because the software packages that can do that are no longer available.

The questions you might ask yourself are:

- Will your designated community just want to be able to download the original file for use by your own systems, in which case no processing is required.

- Will they want to be able to create new information based on the original, in which case you need the material in a high-quality format that is as close as possible to the original.
- Will they want to access the information quickly using widely available software, in which case you need to convert it to a potentially lower quality but more widely supported format.
- Will they expect to access it via an online portal within the preservation system, in which case make sure it is a form that is supported by the built in viewers in the preservation system.

Documenting who these designated communities are, and what they will do reduces risk and helps prepare for future usability.

6. Creating a Format Preservation Policy

Once you know what information you have and how you want to use it, you can create a format preservation policy that ensures it has the best possible chance of being useful when it is needed. The options for each format are as follows:

- Do nothing. This is appropriate if all you want to do is download the file in the future. Just keep the original, make sure the information about it is up to date, but don't convert it to another format.
- Convert to a high quality new "Preservation Master" format that can be used for future content creation or detailed consumption in the future. This is only required if the original format is becoming unsupported or unreliable.
- Convert to a potentially lower quality "Access" format that is widely supported and enables efficient dissemination of the information. This avoids the need for specialist tools and often results in much smaller files that can be accessed quickly.
- Use the Preservation System's built-in render tools to view the information without the need for specialist software. This allows immediate access and is often done in conjunction with one of the conversion policies.

These format policies may have more subtleties. For example, you may choose to apply different rules for information intended for public consumption than you do for internal only information, creating Access copied for public consumption that are not needed internally. You may also want to apply different rules to different parts of your repository, for example having a testbed folder using test rules and the production rules everywhere else.

Given the huge number of formats in the PRONOM format dataset, it can be painstaking and error prone to come up with a rule for every format that might arrive at your system, for example every variant of Microsoft Word. Some preservation systems have predefined rule sets you can choose from that allow you to quickly apply a specific set of actions to a large number of formats with a simple click.

Once created, the format preservation policy should be applied by the preservation system to make sure the content is in step with the policy.

Preservica allows users to create a format rule set that create preservation master formats and access copies, or both, for a huge range of formats. These can be created format by format or by selecting the rule sets prepared and maintained by our Digital Preservation Experts for easy use, for example "Create Access PDFs for all document formats". These rules can be applied to the whole collection, only to certain folders, or to content with specific security settings.

7. Format Migration

7.1 Overview

As discussed above, a user's Digital Preservation Policy may require that files in certain formats are converted to formats more appropriate to their future usage. This should never result in the removal of the original file which should always be kept. However, it could result in several forms of the file being kept for different purposes, all within a single digital asset. This section discussed how these conversions are done and how the asset is maintained.

7.2 Creating a new Preservation Master

A new Preservation Master is required when the original file is not usable using widely supported software, but the information contained is required at as close a quality as the original. The right format depends on the type of information and the context it is being used in.

A widespread example is documents. There are a large number of document formats that are now approaching obsolescence, for example WordPerfect, old Microsoft Word formats and many other obscure formats such as ClarisWorks or StarOffice. If these are needed for editing in the future they need to be converted into a widely supported format, and the obvious choices are proprietary Microsoft 2013 (DOCX) or an open standard OpenDocument Text. Both are legitimate and choosing which to use will depend on whether Microsoft Office is widely used in the organization. Similar choices apply to Spreadsheets and Presentations.

For images the more appropriate target format depends on the original quality. For already compressed, lower quality images it may be appropriate to migrate to Portable Network Graphics (PNG). For higher quality images it is more appropriate to migrate to a less compressed format such as TIFF or JPEG 2000 which can be performed with no image quality loss. Vector images can be converted to the widely support Scalable Vector Graphics format.

Video and audio migrations may be performed to create a new Preservation Master but very often are performed to create access copies. This is discussed below.

7.3 Creating an Access Copy

An Access copy of a file is required where the original format or the Preservation Master format are not widely supported or the file sizes are too large to be easily used. You may also need this where the original format supports editing, but you want to distribute a read-only copy. The file can be migrated to a format that is smaller and more ubiquitous, accepting that in doing so there may be a loss in quality.

The most obvious migration for documents and presentations is to Portable Document Format (PDF). This is widely supported and openly documented and can be read instantly on almost all platforms using freely accessible software. This migration is lossy, that is there is a small loss of quality, for example animations in presentations and hidden text and image quality in documents, but this is often seen as acceptable for the given purpose. Spreadsheets can also be converted to PDF but the loss in quality is much larger – the contents can be read but the sheet is no longer interactive. This may or may not be acceptable, and as discussed above, this may be actively desirable.

Media can be converted to widely adopted streaming formats, MP3 for audio and MP4 for video. Videos should be encoded using an algorithm that supports streaming and can optionally be reduced to a manageable image size to support easy viewing. Raster images can be converted to a widely support compressed format such as JPEG which can be embedded in just about any web page, and vector images can be converted to PDF documents for viewing and printing.

7.4 Migration Tools

There are a number of good quality free tools available to migrate between common file formats, and most digital preservation systems use this extensively. These include:

- LibreOffice – used to migrate between office formats such as documents, spreadsheets and presentations. This converts the incoming format into an internal format then writes out in the new format. Whilst good quality, this change results in some changes and some non-portable features are not taken into the new format.
- FFMPEG – this widely used media transformation tool is used to convert audio and video into new formats. In doing so in can change features such as sample rate or frame size to adjust the quality if required.
- ImageMagick – this is used to convert raster images to other formats.

These tools do a good job with standard formats but struggle with advanced features or some complex formats. One example is the Photoshop set of formats. ImageMagick can convert these, but misses some important features such as adjustment layers meaning the resulting format may be poor quality or worse may reveal hidden information obscured by a layer mask. For situations like this it is better to use the original software, in this case Photoshop, to do the transformation but this is difficult as the software needs to be run on a server in the background and licensing is complex and opaque. The same would be true when using the Microsoft Office suite to transform old Office formats.

Sometimes the software required is itself obsolete. It may be possible to run this in an emulated environment but this is demanding on compute power and again the software licensing is complex and opaque. This is discussed below.

Preservica's preservation tool set combines widely supported tools such as FFMPEG, LibreOffice and ImageMagick with specially built tools to migrate a wide range formats to new formats that are more easily used for the required purpose.

7.5 Ensuring Migration Quality

Ideally it should be possible to algorithmically compare the original and migrated files to confirm the information has not been changed by the format change, or at least that any change is acceptable. For most formats this turns out to be highly complex due to the sheer range of information they can hold and the loose definition of "acceptable".

One file type that can be compared in this way is images. For each image it is possible to calculate an image histogram, plotting the number of pixels for each intensity value, for each of the three primary colors. Whilst some change is expected due to the difference in compression in the file types, any noticeable change can be picked up by allowing only a small variation between the histograms and warning the user it this are exceeded.

Further comparisons are performed by comparing the significant properties of the files before and after. It is fair to say that more work needs to be done in this area to agree which properties are deemed significant and should not change across a migration, and/or the amount of change that is considered acceptable.

Preservica's migration framework supports the definition of "migration validation" actions that can either be configured to do direct significant property comparison between original and migrated files, or call more complex tools, such as the image histogram comparison.

7.6 Handling Change

The IT industry is notorious for its rate of technological change and generally pays little attention to the long tail of information retained for future use. Digital preservation systems have to compensate for this without becoming part of the problem. Whilst a file format policy may seek to migrate files to formats considered to be stable and more widely supported, any policy created today will inevitably be changed in the future and the digital preservation system that implements the policy must allow for this and should be responsive to such changes. The way this can be done is discussed below.

7.7 Maintaining a Digital Asset

As discussed above, when ingesting a file such as myfile.doc, the system may produce a new master format myfile.docx and a new access format myfile.pdf. How this is presented to users depends on the sophistication of the preservation system.

Most, but not all, preservations systems always retain the original file. This is the gold standard information given to the system and given that almost all format transforms lose some fidelity it should always be possible to get to the original, even if this is now difficult to read.

Many preservation systems put the migrated files in a folder alongside the original. This allows all the variants to be read but creates a number of problems. The main problem is that the three files in the example above are only related by name, there is no record of how these files are related, which is the original, and which are the derivatives. A further issue is knowing which of the files has, or should have, the contextual metadata attached, the original or the derivatives? Also, what if you did actually ingest myfile.doc and another unrelated file myfile.docx – how do I know these aren't related?

The best preservation systems create a digital asset within which all the different formats are managed. The user interacts with the asset which has the contextual metadata attached, and the system offers the format that is appropriate to the action that is required. If more migrations are performed in the future, these are done within the same asset. In the example above, the system would create an asset called "myfile" within which the original .doc file, and the new .docx and .pdf files are held. If the user requires an editable version, the .docx is returned but if they want a quick view, the PDF is displayed.

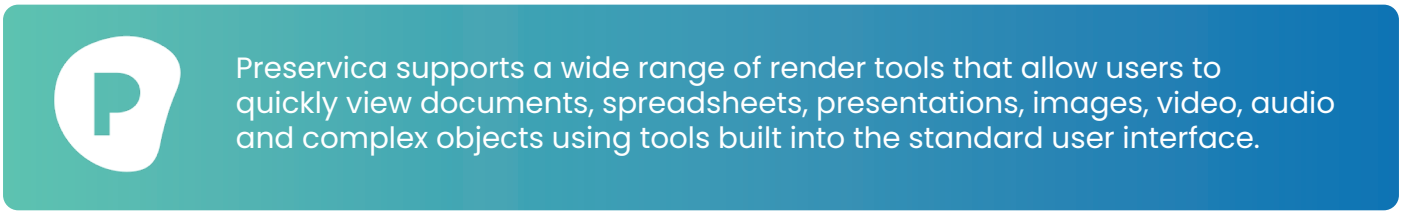
8. Built-in access to the information

8.1 Render

The information inside a file is only accessible using software. This requirement to have access to software that can interpret files of a particular format can make it difficult for users to consume the information. Even after migration to a more widely supported format, it can be difficult for users to install and operate the software needed. Many digital preservation systems embed software, usually called a renderer, to allow the information to be read without the need to install special software.

In many cases, information can be migrated to formats that are easily viewed using built in HTML5 viewers supported by standard web browsers, for example images and videos. These are often wrapped in simple javascript tools to manage the display, for example zoom or moving media forward and back. There are also widely supported tools to display PDF files, and many preservation systems support dynamic transformation of office type formats to PDF for immediate viewing.

Some complex data types require specialist render tools to be built. These include MS SharePoint Records, Emails with attachments, 3D objects, Tweets and captioned videos. This is discussed in the white paper “Preserving complex data”.



8.2 Emulation

In some cases it is not possible to migrate the information to a current format without changing it so much that it cannot be trusted. Also, newer software may create a user experience that is so different from the original that again, the meaning is lost. Lastly, there may not be any tools to migrate the format or any currently supported software that can read it.

In these cases, researchers have built software that simulates old hardware platforms running on existing computing resources. This could, for example, run an old operating system that itself runs an old software package allowing the users to interact with the file in as close a way as the original usage.

Images of these emulated platforms can be kept ready to be deployed as required and can pop up with the operating system and software ready for use. They are useful for documents and presentations but especially useful for interactive information where the way it behaves is critical to its understanding. An example of this is computer games, where the behavior of the game depends on the console being used, but the console is no longer available. Companies such as Electronic Arts create such systems to allow current game designers to be inspired by early versions.

The challenges of using emulation as part of your digital preservation strategy are partly technical. The platforms themselves require significant investment to create due to their huge variety, require significant computing power to host, large amount of storage to be retained, and users require specialist skills to use old software. However the biggest challenges are commercial. Running an old operating system and software package to access information for legitimate research may seem innocent but the use of these systems is protected by international copyright law and it is not obvious whether license fees are due every time they

are used and who is liable for these fees, the user or the host. These issues will need to be resolved before the widespread adoption of emulation as a standard digital preservation approach.

Some Preservica users have built emulation add-ons to their preservation system, allowing users to quickly get access to emulated environments running old software platforms. They maintain a vast library of potential environments, also preserved in Preservica, and allow users to access these, conforming to the copyright laws that apply to their usage. This was descried at the iPres 2018 conference [4]

9. Putting it all together - Automating Format Preservation

Digital preservation systems all allow users to migrate files to new formats. The ease of use and sophistication of this process depends very much on the maturity of the system itself. The options are:

- Automatic migration on Ingest. Most preservation systems allow the user to define a set of preservation rules and then to run these on all content as it arrives at the system. This can be a one-off process meaning that to re-migrate the user has to export the files and re-import, which is generally seen as being highly cumbersome.
- Manual post-ingest actions. Some systems allow the users to manually run a migration process when the preservation policy is updated at a later date. By making this a manual process this can lead to different parts of the dataset being migrated in different ways and human error becomes a factor.
- Full automation: The best preservation systems support full automation. This means whenever the preservation policy is updated the system brings the entire dataset into step with the current preservation policy.

As stated above, Preservica's patent pending "Automated Digital Preservation" capability allows format migration to happen on ingest, or at later date as format policy is changed to match the current user landscape. This happens automatically in the background so all the user has to do is change the policy and the system does the rest. The need for manual intervention is eliminated and the risk of user error is removed.

10. 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. THE TECHNICAL REGISTRY FROM THE NATIONAL ARCHIVES. HTTPS://WWW.NATIONALARCHIVES.GOV.UK/PRONOM/
[2] DIGIPRES COMMONS COMMUNITY-OWNED DIGITAL PRESERVATION RESOURCES https://www.digipres.org/
[3] IPRES 2019: A PRAGMATIC APPLICATION OF PREMIS, O'SULLIVAN, GAIREY, SMITH AND O'FARRELLY, HTTPS://PHAIDRA.UNIVIE.AC.AT/DETAIL/O:1079786
[4] IPRES2017 ADDING EMULATION FUNCTIONALITY TO EXISTING DIGITAL PRESERVATION INFRASTRUCTURE, EUAN COCHRANE, J. TILBURY, OLEG STOBBE, HTTPS://HDL.HANDLE.NET/11353/10.931101

11. Other papers in the Preservica expert series

Digital Preservation Overview

Preserving Multi-Part Information Assets

Digital Preservation Metadata

Digital Preservation Policy Creation

About Preservica

Preservica is transforming the way organizations around the world protect and future-proof critical long-term digital information. Available in the cloud (SaaS) or on-premise, our award-winning Active Digital Preservation™ archiving software has been designed from the ground up to tackle the unique challenges of ensuring digital information remains accessible and trustworthy over decades.

It's a proven solution that's trusted by thousands of businesses, archives, libraries, museums and government organizations around the world, including the UK National Archives, Texas State Library and Archives, MoMA, Yale and HSBC.

preservica.com/about