White paper

Digital Preservation White Paper Series

Preserving Multi-Part Information Assets

Preservica

1. Introduction

Digital Preservation systems are always trying to keep up with the information produced by live content creation packages. As time has gone on, it is increasingly common for this information to require multiple files to accurately represent what you might consider a single piece of information. This paper describes the Digital Preservation techniques required to preserve this type of information and will allow you to compare alternate approaches to solving this complex challenge.

2. What is an Information Asset?

When planning to preserve 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 a single piece of information and in most cases this is a single file, for example a document, image, spreadsheet or video. These simple information assets may still need to be migrated to newer formats and these new formats placed within the same logical asset but each generation of the information can be described by a single file.

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 content files. Removing one of these individual files from these multi-part assets fundamentally changes the information – for example removing one of the attachments from an email means it is no longer the same email.

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". This metadata applies to the whole asset not just individual files.

Preservica manages information as a logical Digital Asset. Depending on the format policy this can create new generations of the preservation masters, and create new access representations for more accessible usage. The user can retrieve any of these and the original, depending on what they want to do with the asset. This asset structure also extends to have multi-file objects as detailed in this paper and was fully described at the iPres 2019 conference. [1]

3. Multi-Part Information Asset Examples

The list of information types that require multiple files to represent them is long and growing as information systems become more sophisticated.

Emails: This ubiquitous information package has evolved over the years. It now comprises an email format file containing the email text (formatted or plain text) plus header information and also a series of attachments. To confuse things further, the attachment can itself be an email file.

Scanned books: When digitized, books, newspapers or magazines are often saved as a series of images, one per page. The order is of course important, so page 1 comes before page 2.

3D objects: Many 3D objects are built from a number of component files specific to the creating software, covering such things as geometry, materials and surface texture. There can also be hierarchies of objects, for example a master geometry file that incorporates a number of sub-components in separate geometry files. The collection of all of these files fully represents the 3D object and they must be preserved together.

Video with multiple captions: Video files are often viewed with caption files to aid accessibility or to allow for translation to other languages. These caption files can also be indexed to make the captioned video findable.

Content management system records: Systems such as Microsoft SharePoint are a way of sharing files but also contain structured data about the file which may be as important as the file itself. Also some records, such as SharePoint Lists can contain multiple content files and some structured data, for example records, can contain complex data types such as images or formatted text. The structured data is much more important than the metadata used as a finding aid and must be preserved within the asset alongside all the content files that between them make up the complete record.

Message board posts: Systems such as Microsoft Teams or Slack allow users to conduct interactions as a series of published posts. Each can contain text as well as attachments, images or media. The metadata about the post, for example who posted it and when and the reactions, may be just as important as the text. The information asset thus comprises the metadata, the text, and any attachments, images and media.

Person to person chat: Personal messages, whether individual or to a group, can be similar to message posts in structure but contain different metadata (who sent the post and when, and which people was it addressed to) along with the attachments and media.

Tweets: Tweets, or X posts, have evolved from simple 140 characters of text, to longer formats of text with a number of media files attached. They also have important status information, for example where the post was sent from and whether it was a reply or re-post. This information can be extracted from an API call or export file.

Web archives: There are a number of approaches to archiving websites, but all of them rely on extracting large numbers of files into a number of containers. It is the sum of all of these web archive files that correctly represents the website.

In many of these cases the meaning of each individual file and which file in the set has primacy must be made clear to enable effective preservation. For example, in a 3D object with a set of geometry files it must be clear which file is the top-level file and which are the sub-components. For an email with another email attached it must be clear which is the master and which is the attachment. These internal structures are specific to the type of asset being preserved.

Preservica supports all of the multi-part asset types listed above and is always adding more to handle the rich and complex data types that are emerging from live information management systems.

4. Preserving Information Assets

4.1 Overview

The preservation of multi-part assets sits on top of the preservation of each file they are built out of and works within a framework described in the paper "Automated File Format Preservation". This allows multi-part assets to be considered a more complex specialization of simple asset preservation.

4.2 Identification

Each individual file component within a multi-part information asset can be identified using the techniques specified in the paper "Automated File Format Preservation". It is the combination and order of these file components that allow us to work out what the multi-part asset is. Here's some worked examples:

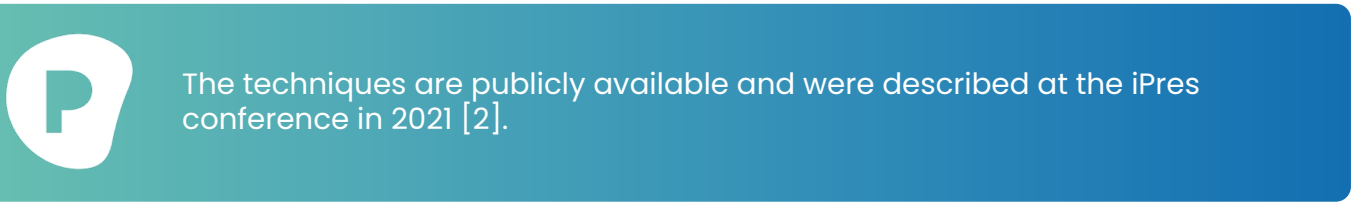
Let's look at an asset that contains a first file identified as a "MIME Email" followed by some office files, for example Microsoft Word 2007 (DOCX) and Microsoft PowerPoint 2007 (PPTX). It is easy to see that this is an email with two attachments. Generalizing this we can say any asset with multiple components and where the first components is an email message format, is an email and attachments.

Another asset may contain a streaming media file, for example MP4, alongside a set of text files identified as "SubRip Subtitle Files" and "Web Video Text Tracks (WebVTT)" formats. This is clearly a captioned video, which can be handled by many HTML5 video streaming tools.

An asset that contains many viewable image files, for example 250 JPEGs, can be considered as a scanned document and can be handled by many widely available book viewing tools.

A more complex asset may contain a set of 3D geometry files, for example in WaveFront OBJ format, plus some material files in Wavefront Material Template Library format, and some JPEG image files. By examining the relationship of the files we can identify this as an OBJ 3D Model which can be rendered as such. A similar pattern can be used for a 3D model built out of GL Transmission Format (GLTF) files and associated formats.

These identification examples are all based on the collection and order of the objects within an asset. A "first object" is often needed, for example a SharePoint database record or an email file. The sequence of objects after this is usually unimportant, although there may be a mandated set of formats that must be present and many that are optional. This list of multi-part asset formats that are recognized by the community is short but growing.



4.3 Validation

Although the identification rules state which formats are allowed to exist in a multi-part asset, often there are interdependencies between the files that must be validated to ensure the asset is usable. The validations usually need special tools to be written that know about the asset type and what is expected. Examples include:

- 3D objects: The initial file, for example the WaveFront OBJ, will list the material files and the sub-component OBJ files that it needs. These lower-level files will themselves contain a list of the files they need, for example the material file lists the texture image files it needs. To validate the asset you need to trace all these internal references and make sure everything is present.
- SharePoint records: These assets contain a database file, in JSON format or similar, which contains the metadata plus a list of the files it expects to be present and why. For example, a SharePoint Document Library item will contain a single attachment and optionally some image metadata files. A SharePoint List Item record may just have the database record file, and optionally can have many attachments, images metadata files and formatted text metadata files. By examining the database file it's possible to validate the contents.
- Tweet (X Posts): Again, tweets (posts) can be extracted as a database file, usually in JSON, that contains the message text plus lots of metadata about the post. It also contains a list of the images and media files referenced by the post, so it is possible to check these images and media files are also present in the asset. The volatile nature of X show these rules need to be dynamic or to be versioned – the number of images and videos and the length of text have all recently changed and further changes can be expected in the future.

All Digital Preservation systems that are able to handle multi-part assets should contain a dynamic list of validation tools to ensure the assets are suitable for future consumption.

4.4 Migration

As discussed in the white paper "Automated File Format Preservation", file formats become obsolete surprisingly quickly and the information may need migrating to allow further high-quality re-use or to create lower quality easy access copies. This same issue applies to multi-part assets as well, with the added complication that some migrations result in the number of files changing, either increasing or decreasing.

The most straightforward approach is to apply the techniques described in "Automated File Format Preservation" to each component within a multi-part asset. For example, if we have an email asset with a MIME Email file and a document attached in WordPerfect 5.1 format, we would expect to migrate the WordPerfect file to both a usable master copy, for example Microsoft Word 2007 (docx) and also create a public access copy in PDF, without needing to migrate the email message file at all.

When requesting a copy of the asset back we can specify which generation to extract and would get those files that comprise that generation, despite the fact that each may have been migrated differently. For example, asking for the access copy we get the original unchanged email file plus the migrated attachment in PDF, but when asking for an editable master we get the unchanged email and the Word 2007 file.

Some migrations are more complex as the number of files in the asset can change. For example, a TIFF image file can contain multiple pages, but when migrating this to create an easily accessible copy it makes sense to create each page as a separate JPEG image file. This increases the number of files contained within the asset but ensures a multi-page viewer could easily present them to a user.

Another example is a Microsoft Outlook Email Message (MSG) file, which can contain its attachments as encoded binaries within the file. This is not ideal as the attachments cannot be individually validated and preserved. A sensible strategy is to migrate the MSG file to MIME Email (EML) and in doing so extract the attachments so the email now only contains the email text and header information. The number of files in the asset will increase by the number of attachments extracted.

It is also possible to go the other way. A scanned publication in a high quality image format such as JPEG2000 can be migrated to compressed images within a single PDF for access purposes. It is also possible to have textual data generated by Optical Character Recognition (OCR) to make the text searchable. The number of files in the preservation representation may be one or two (the image and the text) per page, but in the access representation there is only one file.

Preservica can identify and where required migrate each individual component of an information asset to a supportable format. It can also handle “one to many” and is adding a range of “many to one” migrations, for example a TIFF containing multiple images to many JPEGS and Microsoft Outlook Email to EML with attachments for individual preservation.

4.5 Rendering

Digital Preservation systems contain tools to allow users to interact with assets without the need for specialist software. For multi-part assets these very often have to be built specially, combining widely available tools to present the individual components with a way of presenting the whole information asset as one. This can incorporate components supported natively in the browser, for example video streaming with captions using an HTML5 viewer. It can also combine supported open source components, for example a multi-page book viewer. There are also many JavaScript tools available, for example 3D GLTF viewers.

Whenever tools are used, the Preservation system has to combine them with the logical and physical storage of the information asset so the user has an elegant experience accessing the information. This should include using the most appropriate generation of each component, and the most appropriate representation of the information, bearing in mind any migrations that have occurred. It must also allow all of the files to be downloaded for consumption outside the system.

Preservica incorporates rendering tools for all of the examples listed above. These allow the user to interact with the multi-part asset as a whole as well as downloading some or all of the files they contain for external processing.

5. 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] IPRES 2019: A PRAGMATIC APPLICATION OF PREMIS, O'SULLIVAN, GAIREY, SMITH AND O'FARRELLY, HTTPS://PHAIDRA.UNIVIE.AC.AT/DETAIL/O:1079786
[2] IPRES2021, IDENTIFICATION OF MULTI-PART DIGITAL OBJECTS, O'SULLIVAN, SMITH AND TILBURY, DOI: 10.17605/OSF.IO/YVTS4
https://files.sciconf.cn/upload/file/20211206/20211206163241_26411.pdf

6. Other papers in the Preservica expert series

Digital Preservation Overview

Automated File Format Preservation

Digital Preservation Metadata

Digital Preservation Policy Creation

7. Annex: Multi-part assets compared to traditional paper archiving

Preserving digital information can be similar to preserving physical artifacts. If you were preserving a valuable document from 200 years ago you might seal it in a box which you label and set an access code on to allow only permitted people to see it. Preservica saves digital objects in the same way, saving the document into a logical "digital asset" which you label with metadata and set access rights on.

With the physical document you may decide that very few people can read the old text so you transcribe the words onto a sheet of A4. This is put back in the same box so it has the same label and the same access code. When someone wants to access the document they can ask for a quick view so get the transcribed version. If they want to confirm its contents they are given the original – they contain the same information but serve different purposes, and are saved in the same box.

Preservica treats digital objects in the same way. An old document file such as WordPerfect may be converted into a more modern format such as the latest Word format and a easily readable format such as PDF. These are put back in the same digital asset with the same label and access rights. When a user wants to access the information they can ask for a quick view so get the PDF, for an easily editable copy so get the latest Word, or the original to do some digital forensics. All come out the same logical digital asset with the same metadata and same permissions.

Simplistic preservation systems don't do this. They might create a new Word or PDF but some allow the user to delete the original, which is very risky – imagine destroying your original 200-year old document after doing the transcript. Most also just put all the files in the same folder with a copy of the metadata and access rights. The new files can then be moved, the metadata changed and the access rights updated and they lose their link to the original. Would you just copy the label onto the transcript then allow users to change it and move it – how would you know which transcript belonged to which old document?

At some time in the future it may be found that the transcribed version of the 200 year old paper document is poor quality so a new one is produced. The original is left the same but the old transcribed version is thrown away and replaced with the better version which is again put back in the box.

Preservica does the same with its digital objects. If we find the transformation to the latest Word or PDF was not done well, we can get the original and use better software to re-convert and replace the earlier Word and PDF with the new ones which are put back in the digital asset. This is done automatically at scale for all digital assets in the collection – the next time you ask for a Word or PDF for this asset you get a better quality one.

Simplistic preservation systems can't do this. For a start they might have deleted the original so no re-processing is possible and some systems do not allow migrations after ingest. Even if they have the original and allow post-ingest migration, there are multiple files in the folder so it's not obvious what re-processing you should do and what happens as a result and it's up to you to work that out. Lastly none of this is automated, everything is done at small scale by individual users, there is no collection wide update of all.

Of course some pieces of information need more than one object to describe them. When preserving an old letter you may preserve the letter plus also the photos that were enclosed and also the envelope, which are all put in the same box, with the same label and the same access code.

Preservica treats digital information in the same way. An email for example comprises the email text, the header and any attachments. All are put in the digital asset with some metadata and access control and each can be preserved to make sure they are fit for purpose. When the user asks for the latest version of the email they might get the original email text plus attached documents that have been converted to make them easily readable or can ask for the originals if they want them.

Simplistic preservation systems do none of this. They might allow the user to save all the files in the same folder but there is no obvious relationship between them and the metadata and permissions can be different. Individual attachments can be moved or deleted to invalidate the original email as it is now incomplete or all the objects are kept in one email or zip file making it impossible to preserve each attachment individually.

Preservica's unique advanced Digital Preservation and it's use of flexible and powerful digital assets mimic what we have all been doing for years with physical objects. Combined with full automation, only Preservica can be trusted to make sure you information is as safe and as accessible as any physical object you are caring for.