White paper

Digital Preservation White Paper Series

Digital Preservation Overview

Preservica

1. Introduction

The information which the world depends on has changed. Once exclusively saved on physical media such as paper or film, it is now almost completely digital. This presents huge opportunities – digital information is easier to distribute, can be higher quality and more functional and can be presented in many interesting ways.

However, due to the accelerating rate of technical change this information is under threat, and that threat increases with time. If we think about the technologies we were using twenty years ago, can we imagine what platforms will be in place twenty years into the future? Why would today's information have survived that long and be usable when needed after so many years?

Technological developments also tend to increase the complexity of the digital information we generate; we can see the development from small text files, to formatted documents, to multimedia audiovisual, to 3D and virtual reality objects.

Digital Preservation combines a set of strategies to reduce the risks associated with these threats, aiming to make the information quickly and usefully available long into the future in a way that can be fully trusted. This White Paper and the others in the series show how this growing problem can be managed using sophisticated Digital Preservation technologies to help preserve the world's digital memory.

2. Why Keep Information?

Before deciding to keep information, it is always important to ask why. The reasons for keeping it can be summarized as follows:

- **Legal Protection**: The information may be needed for evidential purposes in the future.
- **Regulatory compliance**: The information is retained to comply with an externally applied set of rules.
- **Repurposing For Value**: The information has inherent value that can be exploited in the future.

The first two are defensive and are about decreasing organizational risk. The third presents an opportunity to exploit the information for benefit at any time long into the future.

3. Digital Information Retention Risk

Whilst holding information in a digital form creates significant advantages, as it is retained for the long term it comes under a wide range of threats that may significantly inhibit the ability to access and trust this information when needed.

- **Constant change**. If the information stays in place in the system that created it, it might be possible to edit it after the decision was made to keep it. Which was the version that was used in an important transaction? This becomes particularly challenging in systems with what is considered ephemeral content such as social media and online publishing.
- **Information backlog**. Most organizations already have a backlog of digital information in aging formats on unreliable storage that is already at risk – this isn't a problem that can be ignored.
- **Permission Loss**. The information may be in place but may become invisible to the users that need to use it. This may be particularly difficult when the information is held by a third party, for example President Trump's Facebook account. [1]
- **Bit level failure**. The 0s and 1s that make up digital information may somehow be changed or damaged in a way that makes the information unusable or untrustworthy. This is a risk

inherent to the physical storage media itself. This risk increases with age and data volume and is especially common in home-cut DVD and CD backups.

- **Storage obsolescence**: The storage media on which the information is held may become unreadable by current technology. A famous example dates from 1986 when the BBC TV show “Blue Peter” saved children’s “digital souvenirs” to a 12” Laser Disk, but when retrieved in 2002 the hardware was unsupported and the software could not be run. [2]
- **Deletion**:The information held on live systems can sometimes be deleted, either accidentally or to free up storage space. If there are no backups it is now lost. Both X (Twitter) [3] and MySpace [4] accidentally deleted old data which is now permanently lost.
- **Format Loss**: Information is formatted to allow it to be interpreted by specific software packages. If the software is not available, cannot run on current systems or the specialist skills to use it are no longer available the file is unreadable.
- **License Loss**: Even if the software is still available, if the license keys are lost the software is unusable. This is especially high risk for hardware keys such as dongles.
- **Context Loss**: If the context of a file is lost then it is difficult to be sure you have the correct version, for the correct purposes and with the correct rights.
- **Attention span**: We live in a post-Google world in which instant access to information is expected. If the information is not findable or can only be accessed using specialist techniques, it might as well not exist.

4. Digital Preservation Standards

A Digital Preservation system is designed to reduce or eliminate as many of these risks as possible. The requirement are documented in ISO 14721:2012 Open Archival Information System (OAIS) [5]. This describes the basics of content management (getting information in, managing it, making it available and extracting it in bulk when exiting the system) along with those processes required to preserve it (bit level preservation, format preservation, and context preservation). The following sections describe each of these in more detail.

5. Inside a Digital Preservation System

5.1 Bit level preservation: durable storage

The first and most basic capability of a DP system is the ability to have complete confidence that the 0s and 1s (bits) that make up digital information are safe, can be returned when needed and have not been tampered with in any way.

The first step is to calculate the checksum for each digital object that the system stores. Checksums are a cryptographic technique that creates a unique text string fingerprint for each digital object - if just one bit in a file changes, so does the checksum. The checksum is stored along with the digital object allowing the system to detect any changes by regular recalculation of the checksum which is compared to the stored value.

When an object change is detected, whether caused by hardware failure, or accidental or malicious manipulation, the system needs to be able to recover. To do this it keeps multiple copies of the digital object on different storage devices. It also makes sure some of these copies are in a geographically remote location. The system can then “self-heal” from local problems like a hard risk failure, or from a catastrophic loss of an entire data center by acquiring one of the other copies.

Preservica maintains the checksums of all objects in the system and checks these at regular intervals to detect for corruption. At the lowest level, it relies on commoditized storage services such as AWS S3 and Glacier and Microsoft Azure Blob Storage which perform all of these actions including geographical dispersal and self-healing out of the box.

5.2 Format Preservation: ensuring usability

Digital information at its most basic level is a stream of bits that is held on a storage device. This is converted into a human understandable form by software, and that software expects the information to be held in a specific way so it can be understood. This pattern is known as a format, which may conform to an agreed standard like an ASCII text file or be proprietary and protected so it can be only read by a specific software package.

The problem is that software packages become obsolete, or at very least not widely adopted, which makes the information in those digital objects difficult or impossible to access. Digital Preservation systems have two strategies for avoiding this problem – either convert the information into a format that can be read by widely adopted software or maintain software that can present the information in the way that it is required. The best systems combine both techniques depending on the intended information consumption.

To put this into action, the first step is to identify the formats the system holds. This is performed by conducting bit level pattern matching using a database of formats such as the PRONOM system maintained by The National Archives, which can be followed by format validation. The properties of the file can then be extracted to support its use in the future.

Now we know the format and some details about the objects, we need to decide what to do to make it usable in the future. We can decide to migrate to a new format to create a new digital master, for example converting WordPerfect documents into the Open Office ODT format. We could also decide to create a lower fidelity but more widely consumed format for access, for example migrating the document to PDF. In fact we can decide to do both, always keeping the original but also creating a new “preservation” copy as a master and an “access” copy for distribution.

The Digital Preservation system may also support tools to allow the information to be presented directly within the system. These “render tools” allow users to interact directly with the information without the need for additional software. This can be simple, for example streaming HTML5 media in a compatible browser. It can also be highly complex, for example simulating obsolete hardware running the original software in a virtual environment.

All of the actions described above are encapsulated in a “digital preservation policy” which dictates which actions are performed on which formats at which times. This can be applied as the information is ingested to ensure the policy is correctly presented and maintained.

The biggest challenge is that the policy will inevitably change, as currently supported formats and software packages change and as our knowledge about historical formats evolves. The best Digital Preservation systems automatically retroactively apply any changes to the “digital preservation policy” or the tools that underpin it to ensure the content is always in step with the policy. This is even more effective when a central team of digital preservation specialists can recommend and distribute the “digital preservation policy” such that non-expert users always have their content preserved using published best practice.

Information should preserved at the lowest atomic level to fully describe a single transaction or piece of knowledge. In most cases this can be represented by a single file, such as an image, audio stream or document. However there are further levels of complexity that arise when information is held in multiple files, for example a Tweet that has the tweet details and attached images or videos. So long as the Digital Preservation system supports these complex objects the same principles apply as with single files.

Preservica delivers the full range of format preservation capabilities out of the box, including format identification, validation and characterisation, and the application of advanced digital preservation policies on simple or multi-part objects. Policies may be set by the user or use the recommendations of our file format experts.

Preservica's unique "Automated Preservation" system allow the automated retroactive application of updated best practice to keep the dataset in step with the current policy.

5.3 Context Preservation: powerful metadata

For a user in the future to fully understand the meaning and status of a digital object, the files on their own are not enough. They will need a set of data attached to the object that tells them more about how and why the object was created, who owns it, and what happened to it before it was preserved. This information, usually described as metadata, can be used to add this context and also to allow for more structured search of the whole dataset to find objects of interest.

The challenge is that each sector and organization has different ideas about which metadata fields are required, and this can vary by data type – what is relevant for audio is not relevant for images. Some standards exist but flexibility is essential to allow each organization to save the metadata they require. The data may be defined as simple name-value pairs that are optionally indexed or complex structured and indexed XML schemas that may confirm to agreed standards. The Digital Preservation system should allow for both and provide tools to load and edit the fields – this is described below.

Some metadata is managed by the system and may be used for system operations. This includes an audit trail of changes to the data within the system, access control information used to dictate who can do what to each object, and structural data to create data hierarchy.

Some metadata is more than just context and gets to the heart of the meaning of the data. This data, often extracted from a records management system, can be written into a data file and added to the digital objects. This data file is preserved in the same way as all content files, using checksum validation to make sure the fields are never changed, ensuring the record data is preserved for as long as required.

Preservica supports customer defined metadata including simple name-value pairs or advanced XML based schemas. Each field can be indexed and a full audit trail is maintained of all changes.

Preservica's unique "Records Preservation" capability allows selected metadata to be preserved at ingest time inside the digital object to allow it to be preserved with full confidence that it cannot be changed.

6. Digital Preservation System Operations

6.1 Acquiring Information to Preserve

As one of the biggest risks to information is leaving it in a dynamic system with no preservation in place, making it easy and if possible automatic to extract and ingest information into a preservation system is critical to an effective digital preservation strategy. This has to be highly flexible to cope with the variety of use cases and include:

- Simple drag and drop or ZIP file upload via a web-based GUI.
- Bulk upload using a secure shared storage location.
- Upload using an Application Programmer Interface (API) from an external software package.
- Automated tools to extract data from their original location and upload them directly into the preservation system.

These approaches should allow for the upload of files, folders and metadata that can be manually prepared before transfer or programmatically extracted and transformed prior to loading.

The ingest process within the system should automatically prepare the information for preservation. This should include basics such as virus checking but also preservation specific actions such as calculating checksums for bit level preservation, the automatic identification of formats, the application of the preservation policy to migrate formats if required, and the indexing of the files and the metadata.

Preservica includes web-based upload as well as load using a shared online storage location to allow quick and efficient ingest. It includes some content acquisition automation, for example website extraction, but where it does not the user can use the API to integrate tools into the system. All uploads can include content files, folders and metadata held in our easy-to-use OPEX format. During ingest the full range of digital preservation preparation actions are applied by the system.

6.2 Managing the Preserved Content

Once the content is loaded, the system should allow the user to change the structure, permissions and metadata. This should be possible using user friendly GUI, via a bulk editor that allows large numbers of objects to be changed in one step, and programmatically using an API. All changes should be noted in an audit trail. These are all standard content management features but still important in a digital preservation context.

The deletion of content in a preservation system is something that needs special consideration given the principal role of the system is to ensure items don't get deleted. By default it should require two people to approve and when this is not possible the system should require special permission to switch this safety feature off. Even after this it should be possible to recover a deletion within a time period of a few months in case of deletion error and all deletions and who approved them should be fully logged.

Note that it is never possible to edit the digital objects themselves. These are saved and have checksums calculated to ensure they are fixed forever.

Preservica allows all users to define, edit, index and export their own metadata fields, whether defined in simple name-value pairs or bulk XML schemas. It allows the bulk edit of metadata fields as well as access and update using the API. Delete by default requires two people and can be reversed for up to 90 days, although this limitations can be relaxed at the customer's risk.

6.3 Making the Information Available

The purpose of a digital preservation system and where it differs from traditional data archiving is that all of this work to preserve information is to allow authorized users to access information quickly and effectively. Users with read permission should be able to browse the hierarchy to find the content they require or to use whole text or fielded search to discover items of interest. It should be possible to refine search results using facets maintained by the system. Users may be internal to the organization with named accounts or external with no account. In both cases access control will dictate which users can see which items in the system.

When a user finds the content they should be able to download the objects for use outside the system or to render them within a browser for quick review. They should also be able to view and download the metadata associated with the content.

Access to the information should be provided by a simple GUI allowing non-experts to find the information. It should also be available via a programming interface (API) that allows alternate GUIs to be used or for the content to be embedded within a third-party system.

Preservica includes an out of the box portal that allows authorised or public users to browse the hierarchy or search for content, within the limits of the specified access control, using whole text, fields or facetted search. Content can be downloaded or rendered using one of the many built in tools.

6.4 System Exit

One of the strange aspects of digital preservation systems is that they should plan for their own demise. To be credible, they must allow the user to extract all of their digital files along with the metadata including the audit trail and permissions and the folder hierarchy to a shared location for transfer to another system in the future. This should be done in as lossless a manner as possible.

Preservica allows users export all content as a hierarchy with full metadata in the human and machine readable Open Preservation Exchange (OPEX) format for re-use in external systems.

7. 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] PRESIDENT TRUMP FACEBOOK ACCOUNT
https://about.fb.com/news/2021/06/facebook-response-to-oversight-board-recommendations-trump/.
[2] BBC DOMESDAY PROJECT
https://en.wikipedia.org/wiki/BBC_Domesday_Project
[3] X (TWITTER) DATA LOSS
https://www.theguardian.com/technology/2023/aug/29/techscape-twitter-x-elon-musk-mass-deletion-images-linkrot
[4] MYSPACE LOSS
https://www.zdnet.com/article/myspace-lost-13-years-worth-of-user-data-after-botched-server-migration/
[5] ISO 14721:2012 OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)
https://www.iso.org/standard/57284.html

8. Other papers in the Preservica expert series

Automated File Format Preservation

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