White paper

Digital Preservation White Paper Series

Digital Preservation Metadata

Preservica

1. What is metadata?

Metadata is defined as “data that provides information about other data”. [1]. In digital preservation the data being described is an asset or a structural folder, and the metadata adds information about that object. There are many different reasons to add information about these objects, relating the digital object management, digital preservation or to facilitate information discovery and exchange.

In digital preservation it is important to distinguish between the preserved asset and the metadata that describes it. The asset itself contains the core essence of the information. It is preserved to a high standard by digital signing and checking and as a result cannot be changed. It is usually in a binary file format that is identified, characterized and may be migrated to a new format. Sometimes multiple files are required to fully define the information.

The metadata attached to a digital presentation asset is usually in a text format held in a database and can be updated as required. Whilst it is important, if it is lost the asset alone still makes complete sense. If the data attached to a digital asset is critical to its understanding and should never be changed it should be placed inside the asset alongside the digital objects it is attached to and preserved to the high standards this implies.

Those working in the cultural heritage sector and records management have a long history of working with metadata for physical assets such as papers, books, images and other artefacts. As information management has moved to a digital form, their experience of metadata management is being put to use in a digital context, and there are many standards and studies of different approaches [2].

Since its first version 20 years ago, Preservica has always allowed users to define and user their own metadata as well as maintaining metadata itself.

2. Why is metadata required for Digital Preservation?

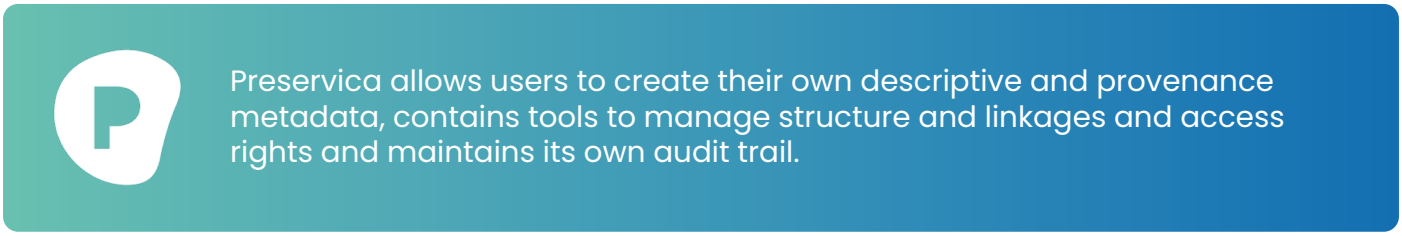
Within digital preservation metadata is created and managed for many reasons, each unique to the manager of the information. This may be to support the business processes of the organization, but may also be to enable the preservation of the objects for the long term.

2.1 Digital Object Management

Metadata may be created and attached to a digital object to understand more about the object or to manage its behavior.

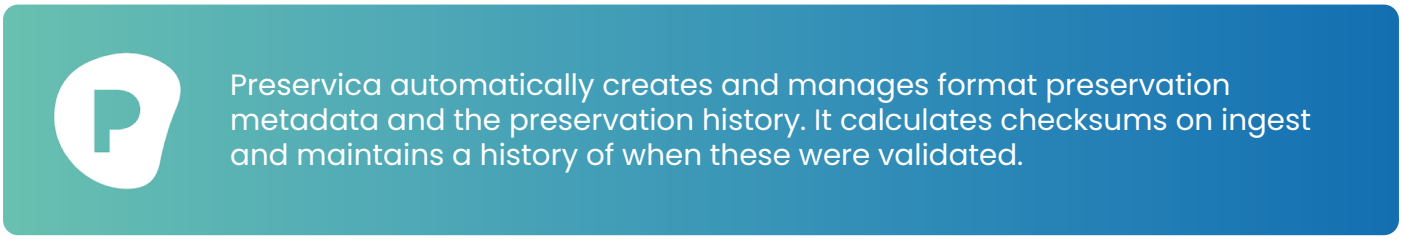
- **Descriptive**: perhaps the most obvious set of metadata is a description of the object to provide context. This may include simple textual description fields or more structured information, for example author, status, reference numbers or usage. This information is often held in standard metadata schemas such as Dublin Core [3] or Encoded Archival Description [4].
- **Structure and Linkages**: Information is very often arranged into a hierarchy of folders that give in context. Links can also be created between assets that are peers, for example “this email is a reply to this other email” or “this document is a later version of this other document”.

- **Provenance and rights**: The history of ownership as well as the current rights are often required to demonstrate what can legally be done with the asset.
- **Access rights**: A list of who can do what to each object is critical to ensure only those people with the correct permissions can create, read, update or delete each object. This should be fine grained enough to ensure each digital object has its own permissions.
- **Audit trail**: A record of everything that has happened to the object within the digital preservation system is essential to demonstrate it can be trusted. It may also be a requirement to retain the audit trail of what happened before the object reached the preservation system, although it is the responsibility of the transfer process to assert this is valid.



2.2 Digital Preservation Technical Metadata

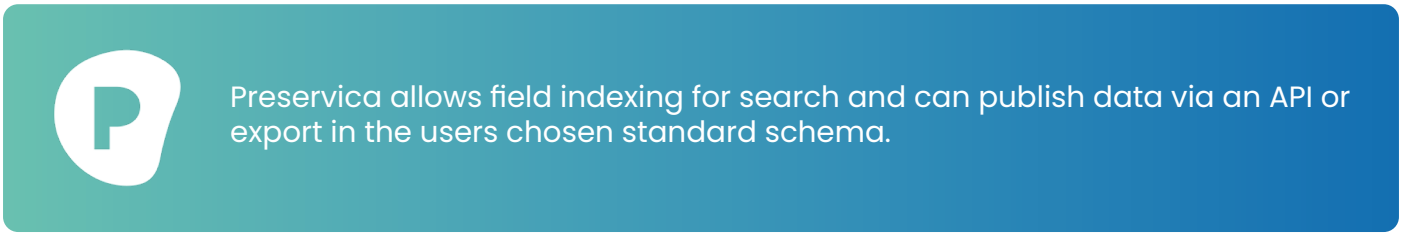
- **Format Preservation Metadata**: The digital preservation system should create and maintain all the information required to ensure the information is held in a format that is usable today. This will include the format of the object, usually a reference to one of the shared format databases such as PRONOM [5]. It will also include technical properties of the objects appropriate to their type to enable any migration processes to ensure the process has been completed correctly. The format and the properties may be updated as the tools and data used to assess them are updated.
- **Preservation History**: A full history of all migrations to a new format or any changes to the technical metadata as a result of improved information about the formats is held in metadata for the object.
- **Physical Preservation Metadata**: Digital objects saved in a preservation system have one or more checksums saved in their metadata [6]. This is calculated by a standard algorithm and can be thought of as a unique "fingerprint" of each file saved. At regular intervals the checksum is recalculated to ensure that the file has not been altered in any way. The checksum and the history of checks are part of the metadata about each object.



2.3 Information discovery and exchange

- **Search**: One of the principal purposes of metadata is to facilitate information discovery. The metadata fields can be thought of as "finding aids" and each can be individually indexed, either as full text words (fielded) or as the whole metadata field (faceted). This fielded and faceted metadata can be used by users to find the correct object via the appropriate user interface. These search fields are often indexed fields created during Digital Object Management.

- **Interoperability**: The transfer of information between systems must include the exchange of its metadata to ensure the new system can fully trust and understand the information transferred.
- **Digital Object Identifier**: DOIs are used to uniquely reference a piece of digital information so a consuming system knows exactly what it is getting. They are often created and maintained by third party organizations.
- **Standards and publication**: Metadata facilitates the automated understanding of the information in an independent manner so it can be exchanged and consumed with confidence.



3. Metadata characteristics

Given all these diverse types and uses of metadata, there are some characteristics that unify all types:

3.1 Dynamic vs fixed

The changes permitted to metadata depend on its usage. It can be considered:

- **User editable**: As metadata is used to support the interpretation of the digital object but is not critical to its use, most of it can be changed in the digital preservation system to allow the user to add context or manage change.
- **System managed**: Many metadata fields are created and managed by the system, including the technical metadata needed for format preservation and the audit trail.
- **Derived and fixed**: Some metadata cannot be changed because it is derived from the object it describes but is extracted into metadata to facilitate search. An example of this is the header fields in an email (To, From, Subject, Data) or EXIF information extracted from an image.
- **Primary and fixed**: Some metadata is critical to the interpretation of the object, for example column values extracted from a content management system such as SharePoint. This may be considered primary data rather than metadata and best practice suggests this should be placed inside the asset and managed alongside the digital objects it is attached to.

3.2 Complexity

The variety of metadata structures is as varied as the objects it describes. There are two principal types:

- **Simple fields**: These are often described as “name-value pairs”, a set of field names and their associated values, which can be created and managed by permitted users. The values can have a variety of types, for example short text, long text, numbers, and dates. Sometimes the values are restricted from a list of values.
- **Advanced Schemas**: Metadata can be more complex, with values that need multiple fields, for example an address, or values arranged in hierarchies. These are often expressed as a standard structure with a set of rules known as a schema, which can be standards across many organizations, and held in a fixed structure in XML [7] or JSON [8].

3.3 Data structure standards

To facilitate interoperability, organisations preserving information have agreed standards for metadata that enable them to exchange information. There are many standards, and these may be applicable to records, archives, libraries, galleries or other organization types. Lists of the standards are maintained in many places including:

ISQ Information Standards Quarterly

Digital Preservation Coalition

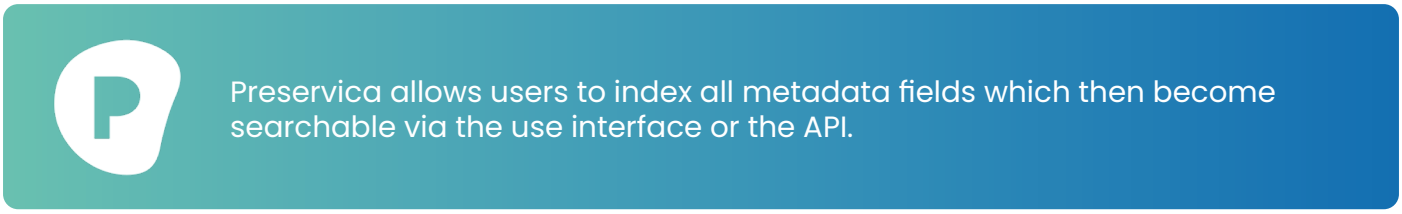
Digital Curation Center

PREMIS

Preservation Metadata (Wikipedia)

3.4 Searchability

As metadata fields are used for search and discovery, they can be indexed in a search tool to allow users to perform advanced search. The type of indexing will depend on the data type, for example allowing numbers and dates to have a range search or an exact match. Text can have either whole field indexing (a facet) or individual words in the field (fielded). This should allow you to distinguish between a search for “find items where the location is New York City” compared to “find items where the location contains the word York”.

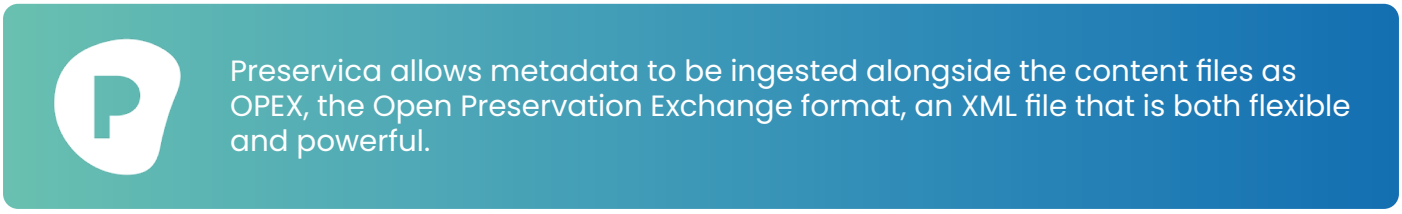


4. Metadata Management

Metadata management is the joint responsibility of the system and the user. Preservation systems provide a wide range of approaches to create, update and export the metadata attached to each object.

4.1 Import

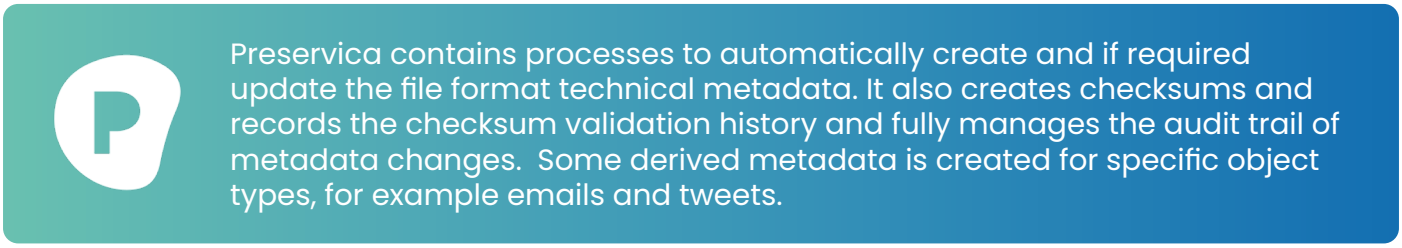
Metadata is often created outside the preservation system by the originating system. It must be possible when ingesting the digital assets into the preservation system to allow its associated metadata to be exported from the original system and ingested alongside the files it described to the preservation system. This allows large scale "export-transform-load" processes to transfer small or huge datasets into the preservation system confident that all the files and metadata are transferred to the preservation system, or the metadata created in a digitization process to be ingested alongside the images it describes.



4.2 Automated creation by the system

Most preservation systems create and if needed update much of the metadata required for file and format preservation and authenticity automatically. This includes:

- **File format technical metadata**: The format information of the files within a preserved digital asset is created and updated by the preservation system.
- **Physical preservation metadata**: The checksums for each file and evidence of their checking will be created and maintained by the system.
- **Audit trail**: All changes to the assets and folders will be maintained by the system, including metadata changes, structure changes and format migrations.
- **Derived metadata**: Some metadata is extracted from the digital asset automatically and put in metadata to allow fielded and faceted search. Examples include email header, EXIF data in an image or document, or posting information for a tweet.



4.3 Automated creation by user processes

Whilst digital preservation systems try to anticipate the automated processes required by users, in many cases there are metadata management processes required that are specific to the business of an organization. Preservation systems allow various extension points to permit user provided software to automatically create or assign metadata to their content including:

- **Built in programs**: Some systems allow users to deploy their own programs into the system workflows, most obviously during ingest, to perform specific actions on the metadata. This can work well for an on-premise system but increasingly this is seen as a security liability and is almost never allowed on cloud hosted or shared systems.
- **Webhooks**: This paradigm is increasingly popular and allows a user defined URL to be triggered after some specific events for a preservation asset in the system. The asset identifier is passed with the URL to allow the program at the other end to use the preservation system programmer interface to create additional metadata for the object.
- **API crawl**: Most preservation systems have a rich programmer interface that allows them to add or update metadata for the assets or folders it manages. This can be used in association with webhooks or as a system crawl checks the entire system and processes the metadata acordingly.
- **Change history API**: These APIs allow programs to ask for all changes since a specific date/time, allowing an external program to catch up with changes and use the other APIs to update the metadata. The most commonly used is the standard OAI-PMH API [9].

These techniques may be used in many ways, for example:

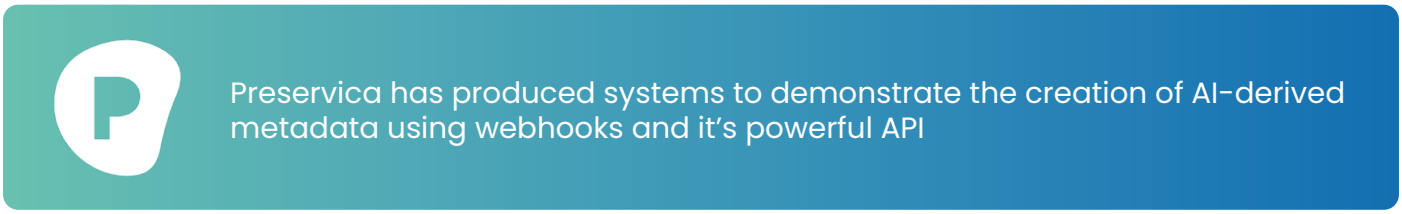
- **Catalogue synchronization**: Every time an asset is ingested or changed in the preservation system, a matching entry in an external master catalogue can be created or updated. The ID of the object in the master catalogue along with other metadata is then written back to the preservation system. This can be two way, so any changes made to the catalogue system can be written into the metadata in the preservation system using an API and a similar model.
- **Creating derived metadata**: As described above, some metadata can be extracted from the digital objects to allow search and discovery. This may be specific to a specific user based on the nature of the objects or some organizational standards. The programmer interface can be used to download the object, process it to extract the metadata and load it back into the system.
- **Metadata field mapping**: Metadata that is automatically created by the system or imported from outside may be in a form that does not conform with organizational standards. User supplied programs can be used to copy or move the ingested metadata fields in the form required and for this updated metadata to be saved back into the preservation system.

Preservica does not allow direct deployment of user provide programs into the system as it is error prone, insecure and breaks the ability of the system to take full responsibility for the longevity of the data. It does support all of the other approaches, providing webhooks for specific events, providing a rich API including the OAI-PHM protocol. There are many examples of users managing their metadata using these tools.

4.4 Metadata creation using Artificial Intelligence or Machine Learning

Both of the system and user automated approaches can be used to link the preservation system to an external AI system which can be used to extract metadata and load it for each digital asset. The uses are potentially hugely useful and could include facial recognition, scene description, personal identifiable information extraction or flagging, document summaries, audio transcription and specialist character recognition. This could result in a metadata update and even automated changes to access control, for example closing an object identified as being sensitive.

This is a new and emerging area so fixed system supplied technology built into the internal processes of a preservation system may be too rigid given the dynamic nature of the domain. Using an external program linked using the techniques described above is a sensible approach given the need to evolve the technology outside of the updates of the core preservation system.



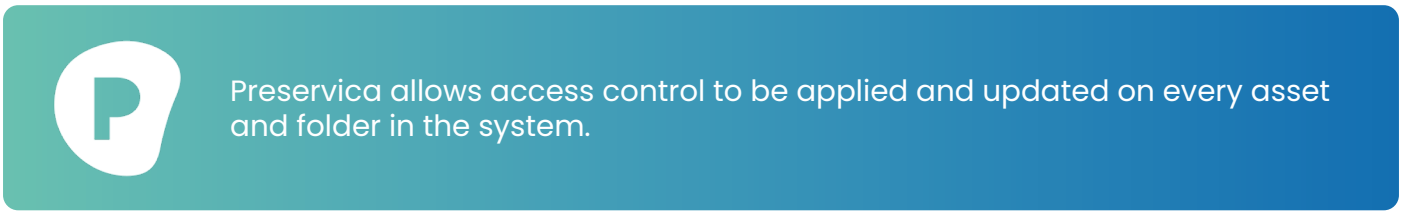
4.5 Manual editing

Most preservation systems allow user managed metadata to be edited using the user interface. This should allow editing of all permitted fields whether held in simple name value pairs or complex schema based structures.

Access to metadata editing should be limited to users with the correct permission.

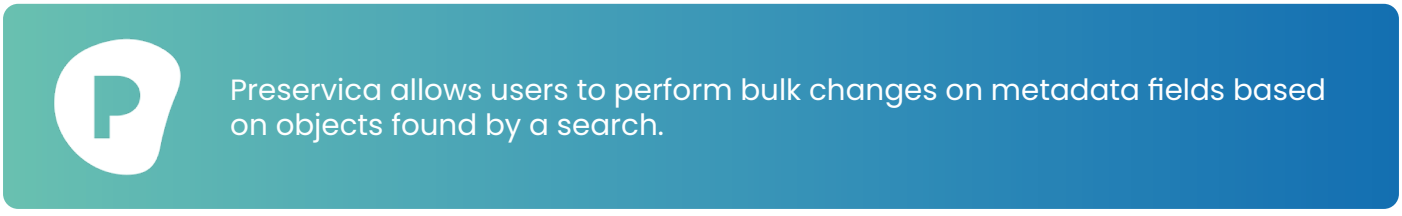
4.6 Access control

Preservation systems should allow users to change the permission on an individual object or a file and folder hierarchy using whatever form they use.



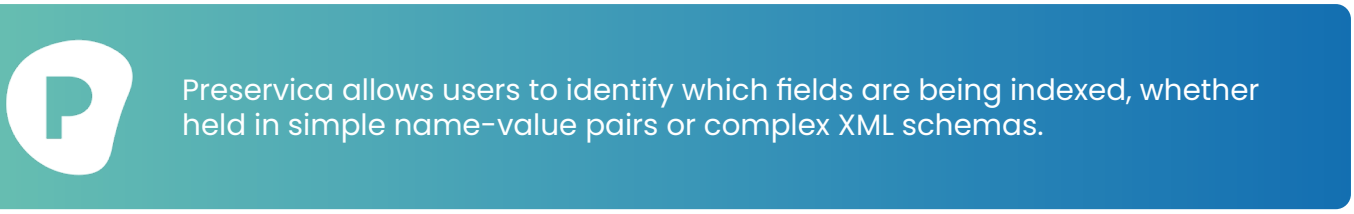
4.7 Bulk update

Preservation systems should allow users to update metadata for a large number of objects at once. This can include making the same change to large numbers or assets or folders, for example a departmental name change. It can also include exporting the metadata for a large number of objects to an external program such as Microsoft Excel, making the changes in the external program, then uploading the data back against the original objects.



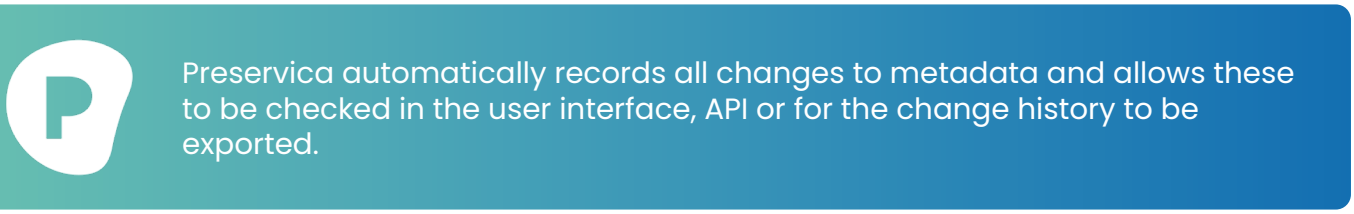
4.8 Indexing

Preservation systems build in the indexing of the required metadata fields to allow the appropriate search technique. This happens when the object is ingested, updated or deleted and is fully automated. The fields to be indexed and the technique used are usually user defined.



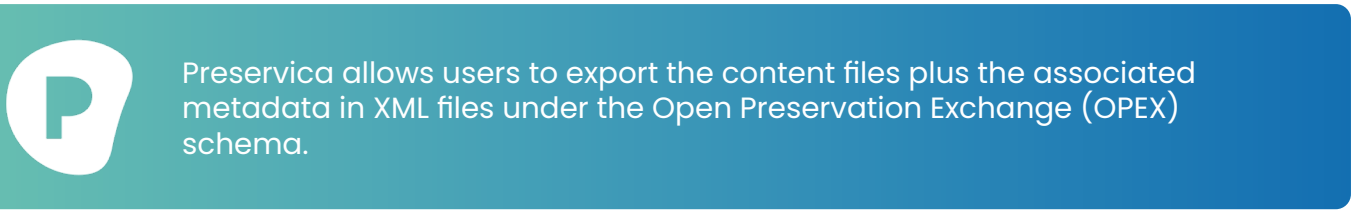
4.9 Change history

The system should maintain a list of all changes to metadata to provide a fully authentic history of everything about the assets and folders in the system. As discussed below this can either be done by holding complete copies of the metadata, one for each change, or holding an audit trail listing just the change made.



4.10 Export

The metadata is important in the life of the objects after export from the system, especially if this is a permanent transfer to another system. A good preservation system will allow the user to export all of the object metadata alongside the object it applies to. This should include all of the metadata types described above.



5. Storing Metadata

The storage of metadata for preservation on computer systems has to balance the reduction of risk alongside the need for performance and efficiency. This balance can seem to pull in different directions, for example reducing risk implies storing the metadata in multiple locations but this reduces performance when making bulk changes to thousands or millions of objects. As such there are often passionately held views on the best approach and so a good understanding of the trade offs is important when making the right choice.

It should also be noted that the files that comprise a digital asset have different constraints to the metadata that is logically attached to them. For example, these files are locked and cannot be changed, they can be huge in size and the files may be held in multiple locations for risk reduction. Metadata is relatively small in size and is frequently updated, often at large scale. Given this it makes sense that the storage approaches for digital objects and metadata can be different.

5.1 Metadata in fixed database columns

Metadata can be held directly in a relational database table with each column corresponding to a metadata field. This is fast and efficient but has a number of very significant downsides. The biggest challenge is that adding or changing the metadata fields is difficult and inefficient. As changes in metadata structure are likely over the length of time of retention this is not recommended for Digital Preservation.

5.2 XML/JSON file separate from the asset files

The metadata can be held in text files formatted using either XML [7] or JSON [8]. The structure can be controlled by a schema that determines the fields held or defined in the data itself as name-value pairs. XML is generally thought of as stricter and JSON as more popular with programmers. There may be a single metadata file per object or multiple, one for each data definition (schema).

There are a number of options for physically saving the metadata files:

File System

Metadata text files can be stored on a regular file system on a shared server. In some cases, they can be stored alongside the content files in a structured manner such that the folder structure on the filesystem matches the logical structure of the digital preservation collection.

This can appear attractive as the metadata is readable outside the digital preservation system and if combined with the folder structure and content files then back door access to the collection is possible if the preservation system is removed.

There are however considerable downsides to this approach. Direct access to the metadata files implies that the access control system can be bypassed and as the preservation system is no longer in control it can no longer guarantee the validity of the information. File system access is also considerably slower when making significant changes, for example a bulk metadata update. Lastly, backup of the file system is delegated to an external system which may or may not do its job making future access to the metadata unreliable.

Object Store: The metadata can be stored in a commercial object store such as AMS S3 which may be done alongside the content files. This adds extra protection as the metadata is stored in multiple locations and check summed. However, it also has similar drawbacks, with slow updates and tempting direct access.

Database: As the text files are small, they can be stored in a relational database system with a unique key corresponding to the object they are attached to. When combined with incremental database backups saving in an object store, this adds extra protection against long term degradation and makes direct access and editing more difficult. This approach particularly excels with very fast access and fast large scale updates.

Hybrid approach: It is possible to store the metadata and indeed the content in multiple locations, having one primary store and one or more alternate stores. The latter can be

managed asynchronously and can be performed using third party developed software using the system APIs to ensure independence. However, it does put a load onto the system and increases the storage requirements and cost.

Preservica saves metadata in XML files in a relational database which is incrementally backed up to an object store. This has been found to be very fast and highly reliable.

5.3 Change control

There are two approaches to metadata change control. The system can either save every iteration of the metadata blocks allowing quick access to all versions of the software. This does however put a burden on the storage to save all the data even the parts that did not change, and is slower to perform at scale.

The alternative is to save individual change records, showing what changed, how and by who. As these are small records they can easily and quickly be stored in a database. This is quick and more efficient, but it is harder to see the state of the metadata at a specific point in time.

Preservica saves the change history as database records, each showing the change, who made it, when and how. This is available via the user interface, the API and can be exported with the object on exit.

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] METADATA DEFINITION & MEANING - MERRIAM-WEBSTER, HTTPS://WWW.MERRIAM-WEBSTER.COM/DICTIONARY/METADATA
[2] DIGITAL PRESERVATION METADATA STANDARDS, HTTPS://WWW.LOC.GOV/STANDARDS/PREMIS/FE_DAPPERT_ENDERS_METADATASTDS_ISQV22NO2.PDF
[3] DUBLIN CORE, HTTPS://WWW.DUBLINCORE.ORG/
[4] ENCODED ARCHIVAL DESCRIPTION, HTTPS://WWW.LOC.GOV/EAD/
[5] PRONOM, HTTPS://WWW.NATIONALARCHIVES.GOV.UK/PRONOM/DEFAULT.ASPX
[6] CHECKSUMS, HTTPS://EN.WIKIPEDIA.ORG/WIKI/CHECKSUM
[7] XML, HTTPS://WWW.W3.ORG/XML/
[8] JSON, HTTPS://WWW.JSON.ORG/
[9] OAI-PMH, HTTPS://WWW.OPENARCHIVES.ORG/PMH/

7. Other papers in the Preservica expert series

Digital Preservation Overview

Digital Preservation Policy Creation

Automated File Format Preservation

Preserving Multi-Part Information Assets