Core Requirements for a Digital Preservation System

Introduction

This document proposes a set of ten high-level functional requirements for a digital preservation system, which is defined as ***the applications and tools used to preserve digital content that enact processes such as ingest, storage, preservation, and access.***

The focus is on the core ***digital preservation*** requirements, rather than the general requirements of an information management system (e.g. security, backup) or the detailed requirements of a particular instance of a preservation system (e.g. number of copies to be maintained). This document is not an attempt to establish a baseline set of requirements for every digital preservation system or context, but rather to capture those features which we believe typically distinguish a digital preservation system from other applications and tools used to manage digital content.

This document forms part of the DPC's Procurement Toolkit which provides advice on different approaches to procuring 3rd party systems and services. It is recommended that this document is utilised in conjunction with the other components of the Toolkit, including the Lessons Learned and Common Requirements.

Benefits

This document aims to simplify and enhance digital preservation system procurement for both the procurer and for 3rd parties responding to procurement exercises[1]. Organizations procuring digital preservation systems may adopt these core requirements as a starting point and then focus on identifying additional functionalities important to them (perhaps specific to their own organizational context, types of content to be managed, existing technological landscape or certification route).

This document may also be used as an educational tool, for example where a practitioner may need to communicate why a typical IT system might not meet the needs of long-term preservation.

Usage

Requirements within this document use "must", "should" and "could" to indicate the weighting of each statement. **Practitioners are encouraged to modify these weightings, the statements themselves and add their own requirements to meet their own needs.** Additional requirements may relate to integration with other systems, workflows for specific content types or additional tasks required of the solution provider (such as data migration). A separate document is available which describes how these requirements can be used alongside DPC's Rapid Assessment Model.

A glossary of digital preservation terms is available in the Digital Preservation Handbook.

Acknowledgements and feedback

These requirements have been created in conjunction with the UK Nuclear Decommissioning Authority. We are also grateful to DPC Members and Supporters who have provided insight and feedback to inform this piece of work. These core digital preservation requirements will continue to evolve in response to feedback whilst remaining closely coupled to the Procurement Toolkit.

[1] Challenges around this process (for all parties) were previously identified in a workshop attended by DPC Members and Supporters. https://www.dpconline.org/events/past-events/guide-to-dp-procurement-event

10 Core Requirements for a Digital Preservation System

1. The system must precisely record and manage the integrity and authenticity of the digital content and metadata it holds. It must use checksums to verify that digital content has not accidentally or maliciously been changed over time and metadata to record any actions enacted on the content.

- Rationale: To verify that digital content remains unchanged over time in a transparent and auditable manner and ensure that planned changes are fully documented.

2. The system must have a comprehensive data model that enables the complex structure of digital objects to be captured on ingest and accurately represented over time as they are managed and preserved.

- Rationale: To enable the system to hold any chosen digital content and maintain the relationships between the elements of complex digital objects so they can be preserved without loss.

3. The system must provide a clear exit strategy to other systems, without vendor lock in. This must ensure that the inevitable migration to a future digital preservation system is possible. It should also minimise the effort and risk involved in such a migration.

- Rationale: To ensure that digital content and all associated metadata can be extracted from the digital preservation system when required.

4. The system must enable authentic digital content and metadata to be ingested.

- Rationale: To ensure content and metadata can be ingested and described without loss or damage

5. The system must have the facility to assess the characteristics of ingested digital content and record them in associated metadata.

- Rationale: Preservation management of digital content is more likely to be possible and successful if the system has information about the nature of that content.

6. The system must support replication and storage management. The system must have the ability to store multiple copies of ingested digital content on different storage systems in different geographical locations.

- Rationale: Effective storage management mitigates the risk of damage to, or loss of content.

7. The system should support preservation planning, including risk assessment, and the design and testing of plans to preserve digital content

- Rationale: To make reliable plans for protecting digital content over the long-term

8. The system should support preservation actions that fulfil preservation plans designed to mitigate identified preservation risks

- Rationale: To enable the steps necessary to protect digital content and ensure ongoing access.

9. The system must support the management of digital content and metadata over time.

- Rationale: Managing and preserving digital content for extended periods is an active process.

10. The system must enable the controlled discovery of, and access to, digital content and metadata.

- Rationale: Digital content is preserved so that it can be found and used by others.

Requirements in detail

1. The system must precisely record and manage the integrity and authenticity of the digital content and metadata it holds. It must use checksums to verify that digital content has not accidentally or maliciously been changed over time and metadata to record any actions enacted on the content.

1.1 The system must record checksums for every file.
1.2 The system must be able to validate checksums against those supplied with content.
1.3 The system must support periodic integrity checking, reporting any damaged or missing files.
1.4 The system must support steps to repair or replace damaged files from replicated copies and report on actions taken.
1.5 The system must be able to generate an audit log and record event metadata (such as that required by PREMIS[2]) describing all actions enacted on digital content.

2. The system must have a comprehensive data model that enables the complex structure of digital objects to be captured on ingest and accurately represented over time as they are managed and preserved.

2.1 The data model must be able to capture and represent digital objects that are composed of multiple hierarchical components, such as: files, drafts, published versions, or copies subsequently created for preservation or access.
2.2 Digital objects should be assigned unique and persistent identifiers.

3. The system must provide a clear exit strategy to other systems, without vendor lock in. This must ensure that the inevitable migration to a future digital preservation system is possible. It should also minimise the effort and risk involved in such a migration.

3.1 The structure of stored digital content could be understandable/interpretable without the preservation system application itself.
3.2 The system must have the ability to batch export digital content and all associated metadata in a manageable format/structure for ingest into another system.
3.3 The preservation system application could be placed in an escrow, providing some reassurance in the event of the failure of the vendor.

4. The system must enable authentic digital content and metadata to be ingested.

4.1 The system must enable the ingest of digital content and associated metadata at scale.
4.2 Content for ingest should be virus checked, with appropriate facilities for quarantine.
4.3 The system must be able to retain not just the original bitstream, but other characteristics necessary for preservation and access to the content such as: original folder structure, file information such as date stamps, associated documentation, and associated metadata.
4.4 The system must be able to provide management reports on the success or failure of ingest activities and should also be able to flag any potential issues with ingested content (e.g. deprecated file formats, unrecognized file formats, missing metadata, password protected or encrypted files etc.).

5. The system must have the facility to assess the characteristics of ingested digital content and record them in associated metadata.

5.1 The system must identify file formats to the level of specific file format version and reference appropriate registries of further information such as PRONOM and/or Wikidata.

[2] https://www.loc.gov/standards/premis/v3/index.html

5.2 The system must extract technical characteristics (such as size, image dimensions, video codec, audio run time, creating application).
5.3 The system should identify content that cannot be rendered, such as broken, badly constructed, or encrypted content.
5.4 The system should validate file formats against file format specifications or customised profiles.
5.5 The system should capture external dependencies, where content (or software) not present in the digital object is vital for it to be rendered or used (such as non-embedded fonts, non-embedded media such as YouTube videos or software libraries).

6. The system must support replication and storage management. The system must have the ability to store multiple copies of ingested digital content on different storage systems in different geographical locations.

6.1 The system must automatically manage the replication of digital content to multiple storage locations (potentially in different geographical locations).
6.2 The system should perform regular system backups.
6.3 The system should be able to regularly test and report on its backup and restore capabilities.
6.4 The system should create and retain management reports on replication, storage management, backup and restore activities.

7. The system should support preservation planning, including risk assessment, and the design and testing of plans to preserve digital content.

7.1 The system should support the identification, management and analysis of preservation risks (e.g. when digital content is in a file format that is no longer supported).
7.2 The system should enable the design, development, and management of plans for mitigating identified preservation risks.
7.3 The system should deliver reporting to enable the effective management of digital content. This must include a wide range of functional, operational, and statistical reports and analytics.

8. The system should enable preservation actions that fulfil preservation plans designed to mitigate identified preservation risks.

8.1 The system should enable the migration of files from one file format to another.
8.2 The system should enable the rendering of digital content via the application of emulation and/or other specialist tools.
8.3 The system should quality assure the results of any preservation action.
8.4 The system must record preservation actions (and their outcomes) in the associated metadata.

9. The system must support the management of digital content and metadata over time.

9.1 The system must provide controls to minimise the risk of accidental or malicious deletion or content change.
9.2 The system must be able to ingest and manage all required metadata including metadata appropriate for specific content types (e.g. geospatial, audio visual).
9.3 The system must facilitate the enhancement of metadata and documentation throughout the lifetime of the digital content.
9.4 The system must enable the managed disposal of content and permit the retention and maintenance of metadata even when the associated digital content has been removed from the system.
9.5 The system must restrict any actions to manage digital content, based on configurable user roles.

10. The system must enable the controlled discovery of, and access to, digital content and metadata.

10.1 The system must ensure that all digital content and any associated metadata is only discoverable and accessible by authorized users.

10.2 The system could warn users if they attempt to access any content that may be problematic (e.g. digital content which is held in an unsupported file format, has incomplete metadata, requires additional permissions etc.).

10.3 The system could provide emulation facilities, specialist viewers, migration on-the-fly, or helper applications to enable the use of digital content in obsolete or unsupported formats by users.

10.4 The system must provide an access interface for users and/or sufficiently capable API to enable integration with other user access systems or discovery tools.

10.5 The system should expose information about digital content on access, including preservation metadata and checksums, to demonstrate a chain of authenticity.

10.6 Metadata and/or digital content should be harvestable.