Digital preservation requirements for procuring IT systems

Introduction

This document proposes the requirements that should be considered when procuring an IT system (for example, an EDRMS, DAMS or GIS), that may ultimately contain at least some records or digital content that need to be retained beyond the life of the system. It must therefore be possible to select and extract content and metadata for ongoing management and preservation.

This document forms part of the DPC's Procurement Toolkit which provides advice on different approaches to procuring third party systems and services including a related set of core requirements for procuring a system with the primary purpose of preserving digital content.

Benefits

By enabling content extraction in a managed way from IT Systems, organizations can avoid costly barriers to preservation and/or migration of content when an IT system is retired at the end of its life and ensure continued access to valuable information.

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 include additional requirements to meet their own needs.** We anticipate that this document will be used by digital preservation practitioners and information and records managers when contributing requirements for the procurement of IT systems within their organization.

Acknowledgements and feedback

These requirements were created in conjunction with the UK Nuclear Decommissioning Authority. We are also grateful to DPC Members who have provided feedback. This document will continue to evolve in response to feedback whilst remaining closely coupled to the Procurement Toolkit.

Principles

The following principles should be applied to ensure that any content ultimately selected for long-term preservation is managed effectively before transfer to a digital preservation system:

- Appropriate records management policy and procedure should be put in place and fully documented, including clear criteria and related processes for record disposal, retention, and long-term preservation.
- Data and metadata should be structured in a way that makes it straightforward to use and re-use beyond the life of any particular IT system. Open data standards, metadata standards and file formats that facilitate data interoperability are encouraged.
- Robust processes for backing up current data should be applied.

Acceptance Testing

It is essential to verify that a product selected during a procurement process does in fact meet the specified requirements in practice. The inclusion of a content import, extraction and preservation scenario as part of user acceptance testing is therefore recommended. Testing with a sample set of content and metadata that has been extracted from the IT system can be useful in flagging up any issues before it is too late to make significant changes.

Requirements

The following are critical requirements for consideration in a procurement exercise:

1. The system should use appropriate open data standards to structure and store data.

- Rationale: Data standards facilitate subsequent data interchange and interoperability without the need for costly and/or complex data migration.

2. It must be possible to import and store content and associated metadata, if the system is to be populated with existing data.

- Rationale: If the system to be procured will initially be populated with data from an existing system that it is replacing then it will be necessary to ensure that the data as well as accompanying metadata can be effectively imported and stored.

3. The system should enable digital content to be selected for disposal or retention/preservation as appropriate:
   a. Flagging of content by users for action or for specific retention periods.
   b. Selecting content for extraction using search on content and/or metadata.

- Rationale: Not all content held within a system will be of equal value or will need to be kept for the same period of time. Being able to manage retention periods and mark content for deletion or for preservation are important features to help ensure that the right content is managed for the right period of time.

4. The system must provide a practical mechanism for the extraction of digital content, such as via an API and/or user interface.

- Rationale: An IT system has a finite lifespan. Suitable export options must be available if the content held within the system has a retention need beyond the life of the system itself.

5. The system must enable appropriate metadata, structural and contextual information to be extracted along with the digital content. This could include:
   a. Descriptive metadata.
   b. Structural relationships between files or other items of digital content.
   c. Fixity information.
   d. Datestamps.
   e. Unique identifiers.
   f. Audit logs (for example who created and edited the content and associated dates).

- Rationale: Digital content may be of little value without metadata that helps it to be located, understood and trusted.

6. The system must allow the extraction of digital content and metadata in formats that will permit its use outside of the system.

- Rationale: Dependence on an obsolete IT system may hamper or prevent the understanding and use of digital content and metadata.

Examples

The examples that follow illustrate a range of IT systems and scenarios where consideration and testing of these requirements has occurred, leading to better outcomes for the longevity of the digital content.

Example 1: Records in a Library Management System

Bibliographic records held in a library management system can be selected and exported in the MARC 21 format – which has been widely adopted by the international library community, and is well-supported by library management systems. Most current library management systems have robust mechanisms to ensure the successful export and import of MARC 21 bibliographic data. By comparison, transactional data (e.g. records of which books have been borrowed) is much more challenging to share reliably between library management systems because the NISO Circulation Interchange Protocol standard (ANSI/NISO Z39.83) has not been so widely adopted. This has led to the use of bespoke export/import and conversion solutions which can result in changed or lost data. Acceptance testing included a focus on extraction of this transactional data to ensure that the export routines produced data that was fit for purpose.

Example 2: Images held in a Digital Asset Management System (DAMS)

Images and associated descriptive metadata are to be held in a new DAMS which replaces an old system which has reached end of life. Initial testing is carried out to ensure that existing content and metadata from the legacy system can be successfully imported into the new DAMS with no unacceptable information loss. Further tests are carried out to assess whether export capabilities are fit for purpose. It is possible to perform checks to establish that selected image data (stored as TIFF 6.0) can be successfully extracted from the DAMS system without any data loss. Additional tests were carried out to establish if essential metadata (e.g. descriptive information, identifiers, and checksums required to establish file integrity) could also be reliably selected and exported.

Example 3: Video managed in a Content Management Systems (CMS)

Digital content produced by a new media company is stored, managed and accessed from a CMS on a day-to-day basis. A search facility allows specific titles to be selected and extracted for long-term preservation. An extraction facility allows these selected videos to be exported from the CMS. Along with the video files themselves, basic descriptive metadata is also extracted along with an internal checksum used by the system and a unique identifier which is present in the filenames of the exported content and metadata files, ensuring that the link is not lost between digital content and its associated metadata.

Example 4: Records held within an Electronic Document and Records Management System (EDRMS)

Documents are held in an EDRMS which is widely used by staff within an organization to store and access current records. Periodically, records which are considered to be of permanent value are extracted from the EDRMS and transferred to a relevant archival authority. As well as the records themselves, it is essential that associated metadata is also transferred. Export tests carried out have demonstrated that descriptive metadata assigned to a record within the EDRMS is exported in a suitable format, but highlighted that additional work was required to ensure that the last modified date of the file itself was properly captured. By default, all records extracted from the system are assigned new system dates and thus lose valuable information relating to when the record was created and last modified. Further tests were carried out to ensure that information relating to creation and edit dates is captured in metadata and can be extracted alongside the records themselves. Given that relationships between records within the EDRMS are also an important characteristic to preserve, checks were also made to ensure that local system identifiers used to create relationships between different records were also included in the export.