Digital Curation 101

INGEST

Ingest is the fourth sequential stage in the curation lifecycle, following *Appraise & Select*.

Topics:

- The ingest stage
- OAIS and ingest
    - SIPs, AIPs, and PDIs
    - METS (Metadata Encoding and Transmission Standard)
- Ingest processes in more detail
- Ingest tools
- Policies

The ingest stage

Ingest is the fourth sequential action of the data curation lifecycle. Its activities are:

- Transfer data to an archive, repository, data centre or other custodian
- Adhere to documented guidance, policies or legal requirements.

The term ingest refers to the process of adding objects to a preservation repository. Ingesting materials into a managed repository environment is a prerequisite for effective curation. Note that the data or digital objects to be ingested have already been appraised and selected as appropriate for long-term curation.

Ingest refers, broadly speaking, to the processes undertaken when data is newly accessioned and before it is added to the main repository. The activities covered by the *Ingest* action include:

- Receiving appraised data
- Preparing them for placing in long-term storage, such as a preservation repository. This involves:
    - Assigning a persistent identifier to the data
    - Checking that the data does not contain malicious spyware or malware
    - Extracting, creating and assigning relevant description and representation information
    - Creating fixity values (i.e., digital signature, hash value) to assist checking the integrity of data

  - Confirming technical details (i.e., file format, MIME type)
  - Combining the data and their associated description and representation information into an Archival Information Package (AIP)
- Possibly, migrating data to a different file format

OAIS and ingest

The term ingest and its associated actions are derived from the OAIS reference model. Ingest is the first of seven functions outlined by OAIS. A key concept in OAIS is the Information Package (IP), which consists of:

- The digital object to be preserved
- The metadata (description and representation information) required at that point in the system
- The Packaging Information linking the digital object and the metadata.

Two kinds of Information Package, the Submission Information Package (SIP) and the Archival Information Package (AIP), are important for the ingest action.

SIPs, AIPs, and PDIs

A SIP comprises the digital object and its accompanying metadata as presented at the start of the ingest action.

An AIP is based on a SIP, to which additional information needed to manage preservation Preservation Description Information (PDI) is added. A PDI has four components:

- Reference Information - a unique and persistent identifier
- Provenance Information - the history of the archived object
- Context Information - relationship to other objects, for example, the hierarchical structure of a digital archive
- Fixity Information - a demonstration of authenticity, such as a hash value.

Also maintained should be the representation information required to render the object intelligible to its designated community. For example, information regarding the hardware and software environment needed to view the content data object should be maintained.

Digital Curation 101

Metadata Encoding and Transmission Standard (METS)

The METS standard is used to provide the Packaging Information by associating all the metadata about a digital object, including that object's relationships with other objects, with the object.

Ingest processes in more detail

The ingest action consists of the following processes. Not all of these processes may be needed: for example, a digital object may not be encrypted and therefore unencryption will not be required. This section is based on material developed by the Complex Archive Ingest for Repository Objects (CAIRO)[1] project.

- Assigning a persistent identifier identifier

- Providing metadata, both technical and descriptive. Technical metadata records information about technical characteristics of the data such as file format, and often is automatically extracted from the data. Descriptive metadata records information about characteristics of the data, such as where created and who created it, and is usually assigned
    - Extracting metadata:
        - Extracting general metadata embedded in a digital object and common to all formats, for example, file size
        - Extracting metadata specific to an object type and embedded in the object, for example, text or image
        - Extracting metadata specific to a format type and embedded in an object, for example, TIFF header metadata
    - Creating metadata:
        - Reading file/system information or incorporating metadata added manually
        - Creating a digital signature
    - Normalising metadata: converting it to another form, for example, normalising dates to the ISO Standard
    - Validating metadata in XML to ensure it conforms to appropriate XML schemas

- Adding fixity information that guarantees that files have not been changed during

[1] http://cairo.paradigm.ac.uk/

storage, for example, by generating a checksum during the ingest process which is then verified in later processes:

  - Generating a checksum, usually using freely available software
  - Generating a digital signature

- Confirming technical characteristics:
  - Confirming file format, using JHOVE ID, DROID ID or other tools
  - Confirming MIME type

- Checking for the presence of malware

- Creating a wrapper or container into which metadata can be placed, for example, the creation of the METS AIP file containing XML metadata

- Migrating an object to a different file format

- Uncompressing digital objects received in compressed formats

- Unencrypting digital objects received in encrypted form.

Ingest tools

Automating the ingest process is currently limited by the lack of tools to assist in the ingest workflow. Stand-alone tools exist for some ingest processes, for example, generating technical metadata and creating checksums.

Examples of ingest tools in common use include:

- Digital Record Object Identification (DROID)[2] performs automated batch identification of file formats
- Jacksum[3] (a tool for calculating and verifying checksums, hash algorithms and Cyclical Redundancy Checks (CRCs)
- JSTOR/Harvard Object Validation Environment (JHOVE)[4] performs format identification, validation, and characterisation of digital objects

[2] http://sourceforge.net/projects/droid/
[3] http://sourceforge.net/projects/jacksum/
[4] http://hul.harvard.edu/jhove/

- NLNZ Preservation Metadata Extraction Tool[5]: automatically extracts preservation-related metadata from the headers of a range of file formats, and outputs that metadata in XML.

A fuller listing of ingest tools is provided in the *Cairo Tools Survey: A Survey of Tools Applicable to the Preparation of Digital Archives for Ingest into a Preservation Repository*[6]. The CAIRO project is developing an integrated ingest too. The aim is to bring together existing tools into an integrated automated workflow that produces metadata packages in the form of METS files.

Policies

As is the case with all actions in the curation lifecycle, *Ingest* is best implemented if policy statements and guidance are developed. These policies need to be documented and kept up to date, for example, with respect to changes in legal requirements.

The reasons why documented policies are useful for ingest procedures include:

- They clarify responsibilities and lines of communication
- They promote standardisation
- They allow risk management
- They ensure IPR and compliance issues are addressed.

For example, an ingest policy could provide guidance for and specify the repository's policy about submission and file formats, addressing questions such as:

- Does the repository have a policy on submission file formats?
- Are there any restrictions on files formats ingested?
- Does the repository transform submitted formats in any way?

Examples of documented policies for ingest procedures include:

- The British Library of Political and Economic Science's policy on versioning[7]
- Yale University and Tufts University's Ingest Guide for University Electronic Records[8].

[5] http://meta-extractor.sourceforge.net/
[6] http://cairo.paradigm.ac.uk/projectdocs/index.html
[7] http://www.lse.ac.uk/library/vif/Framework/RepositoryManagement/policies.html
[8] http://dca.lib.tufts.edu/features/nhprc/reports/3_1_draftpublic2.pdf

Digital Curation 101

The next action in the curation lifecycle

The next sequential action in the curation lifecycle is ***Preservation Action*** which investigates actions to ensure long-term preservation and retention of the authoritative nature of data.