LSHTM DATA COMPASS:
DIGITAL PRESERVATION PROCEDURE

Document Type	Procedure
Document owner	Gareth Knight (Research Data Manager)
Approved by	LAS Management Team
Approval date	March 2019
Review date	March 2021
Version	2.0
Amendments	Policy changed to procedure
Related Policies & Procedures	See References section

1. SCOPE (i.e. who does this affect)

This procedure is relevant to LSHTM researchers who wish to understand the process that will be applied to curate and preserve digital objects in LSHTM Data Compass, as well as LSHTM Library & Archives Service staff who will implement the procedure.

2. PURPOSE

LSHTM Data Compass is a digital repository of research data and other digital objects produced by researchers at the London School of Hygiene & Tropical Medicine and its affiliates

This procedure outlines processes that will be applied to manage digital objects that the LSHTM Library & Archives has agreed to curate and preserve for the purpose of making available through LSHTM Data Compass. Specifically, it sets out how it will address three objectives:

1. Protect the integrity of digital objects by monitoring for evidence of unauthorised modification or deletion and taking corrective action (bit preservation)

2. Ensure information content and other significant properties of the digital object remain accessible across changing technological environments (content preservation)

3. Enable information content to be used in scientific research by providing information sufficient to access, understand and use it (documentation)

The Procedures builds upon and addresses the requirements of several standards, including the OAIS Reference Model, Concordat on Open Research Data, and Core Trust Seal.

3. PRESERVATION STRATEGY

To achieve the three objectives outlined, the LSHTM Library & Archives Service has adopted a preservation strategy built upon a risk management approach. This consists of three components:

1. *Prevent:* Apply measures to reduce the likelihood that a problem will occur in the future.

2. *Monitor:* Monitor for indicators that a problem is developing or has occurred.

3. *Correct:* Assess the problem and take action to rectify or reduce its severity.

3.1. OBJECTIVE 1: PROTECT THE INTEGRITY OF DIGITAL OBJECTS

Digital object integrity is based upon the notion that it has not been altered as a result of an unauthorised event, resulting in one or more files being modified or deleted. Events that may lead to integrity loss: deliberate changes by an unauthorised user, file corruption as a result of gradual media failure or poorly-configured software tool, and modification caused by malicious code (such as a virus). The following steps will be taken to manage these risks:

PREVENT:

a. Ask data providers to ensure each digital object is complete and free from corruption prior to submission. Work with the data provider to obtain a corrected copy.

b. Backup digital objects at regular intervals (daily) to two or more storage media, one of which is held in a different location.

c. Apply access controls that limit the ability for people and processes to modify or delete digital objects. Data providers should only be able to edit digital objects in their work area; they should not have permission to modify files after they have been deposited and made available.

MONITOR:

d. Generate a baseline checksum / hash value for each digital object as soon as it has been submitted to the digital repository. The chosen checksum algorithm should be collision-resistant.

e. Perform a fixity check at regular intervals (e.g. daily) by generating a checksum / hash value and comparing it to the baseline value. The fixity check should be performed upon the production server and any backups.

f. Monitor people and processes that have write access to digital objects and restrict access when access is no longer required, e.g. a staff member leaves.

CORRECT:

g. Identify the problem that has caused digital objects to be modified or deleted and take action to prevent it happening again, e.g. remove the virus, disable the poorly-configured script, disable the user account of a malicious user.

h. Identify digital objects that have been modified/deleted and restore these items from backups at the earliest opportunity.

3.2. OBJECTIVE 2: ENSURE SIGNIFICANT PROPERTIES OF THE DIGITAL OBJECT REMAIN ACCESSIBLE

The ability to access and analyse the information content contained within a digital object is dependent upon the availability of software tools capable of interpreting the encoding format and rendering it in a form usable to the researcher. Several scenarios in which a digital object can become inaccessible or unusable may be recognised. These include: the file format is only supported by one specialised software tool that is no longer sold; the latest version of the software does not support functionality required by the digital object, implements it in a way that is incompatible, or implements it in a way that results in unexpected changes; software tools that support the file format apply different interpretations of the specification, resulting in expected changes; or the researcher cannot afford to pay for a software licence.

To ensure information content and other significant properties of a digital object are accessible to the widest user community, a strategy of active intervention through format conversion is adopted for most types of research data. Alternative strategies such as emulation will be considered on a case-by-case basis.

PREVENT:

a. Produce and publish guidance on file formats that should be used when submitting digital objects (recommended, acceptable, non-acceptable). The use of open, well documented formats will be encouraged, following recommendations by the UK Data Service, The National Archives UK, and other experts.

b. Ask data providers to provide digital objects in a machine readable, open format that can be interpreted by a large number of software tools (e.g. CSV, tab-delimited text). If this is not feasible, data providers should be asked to specify the significant properties to be maintained and provide digital objects in a form that can be accessed using software tools available to LAS staff. Preservation action to export significant properties should be performed at the earliest opportunity after it has been received (known as a migration-on-ingest strategy).

c. Liaise with the data provider to obtain information on encoding formats and software tools at the earliest opportunity.

d. Ensure appropriate permissions are obtained from rights holders to enable LAS staff to perform preservation action.

e. Undertake a Technology Watch each year to identify signs of emergent issues, e.g. file format obsolescence, change in the capabilities of the user community[1], change in financial costs for processing.

f. Produce and maintain a Preservation Plan for each digital object class. The plan should indicate significant properties to be maintained, preferred file formats for preservation and distribution, recommended conversion paths (including use of specific tools), and checks to be performed.

MONITOR:

g. Apply characterisation tools capable of determining the encoding format and identifying characteristics at the earliest opportunity. Store characterisation results as technical metadata (e.g. in XML format) for review

CORRECT:

h. In the event of risk factors being recognised in one or more digital objects, a preservation plan outlining how the problem will be addressed should be developed and submitted for approval by the LAS management board.

i. Perform preservation action and record an audit trail. The audit trail should describe the task performed, software tools/processing script applied, when it was performed, the name of the person who oversaw the process, and other relevant details.

j. Update public metadata record. A digital object that has been exported to a new preservation format can be added to an existing metadata record without creating a new version on condition that no content changes have been made. The obsolete format should continue to

[1] The user community equates to an OAIS Designated Community – the recognised target audience of the research data repository who should be to understand the information provided.

be made available if it can potentially be useful to some researchers and there is no risk associated with continuing to make it available.

3.3. OBJECTIVE 3: ENABLE INFORMATION CONTENT TO BE USED FOR SCIENTIFIC RESEARCH

The ability to understand and use digital objects for new research is dependent upon the availability of supporting documentation. A digital object that does not possess sufficient documentation is unlikely to be used and may have less research impact as a result.

To ensure sufficient contextual information is provided to enable digital objects to be accessed, analysed, and applied in further research, LAS staff will check documentation and work with data providers to ensure it is of sufficient quality.

PREVENT:

a. Provide guidance and a template that outlines the type of information to be provided with each digital object.

MONITOR:

b. Review documentation provided by data providers and note gaps and discrepancies.

CORRECT:

c. Produce updated documentation with input from the data provider.

4. DEFINITIONS

- *Data contact:* One or more people responsible for making decisions with regards to data collection. For example, the Principal Investigator of a project in which data was generated.

- *Digital object:* A unit of digital content that consists of one or more files, including research data, metadata and other files (https://definedterm.com/digital_object/127687).

- *Data provider:* A person or organisation that has provided a Digital Object to the research data repository for curation and preservation.

- *Digital repository:* A system designed to store, manage and potentially make available digital objects.

- *Metadata:* Data that describes the characteristics of other data, such as spreadsheets, databases, and other content types.

- *Research data repository:* A type of digital repository that primarily handles research data.

5. CONTACTS

Questions related to the Digital Preservation Procedure and its implementation should be directed to the Research Data Manager based within the Library & Archives Service (researchdatamanagement@lshtm.ac.uk).

6. REFERENCES

- Concordat on Open Research Data http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/

- Core Trust Seal https://www.coretrustseal.org/

- LSHTM Research Data Management Policy
  https://doi.org/10.17037/PUBS.00612422

- LSHTM RDM Policy support document - Data Access Procedures
  https://doi.org/10.17037/PUBS.00612422

- LSHTM Records Retention & Disposal Schedule
  https://lshtm.sharepoint.com/Services/Information-Management/Pages/-records-retention-and-disposal-schedule.aspx

- Reference Model for an Open Archival Information System (OAIS). Recommended Practice, CCSDS (Magenta Book) Issue 2, June 2012
  http://public.ccsds.org/publications/archive/650x0m2.pdf