Basic Principles of Digital Archiving

1.0 Introduction

This paper sets out the underlying principles of digital archiving. It does not give specific recommendations for formats, hardware, or applications, but instead sets out the fundamental elements required by a digital archive, so that archives are able to assess the suitability of specific solutions and approaches, and plan the resources and funding needed to set up and sustain their systems.

2.0 Digital Preservation Standards

In order to comply with international best practice for the long-term management of digital assets, an archival preservation system needs to fulfil a number of important requirements.

Although there is no universally adopted standard for digital preservation, the *Open Archival Information System (OAIS), ISO 14721:2012* is the most widely adopted model. The OAIS has its roots in the space industry's need to preserve large amounts of digital data and, although daunting to read in its entirety, its essential points can be easily summarised. OAIS is linked to another standard, *ISO 16363: Space data and information transfer systems – Audit and certification of trustworthy digital repositories*. This is intended to provide a standardised method of auditing a digital repository, and an archive which satisfies all the criteria can be formally granted 'trustworthy' status. It is perhaps the most useful aspect of the OAIS model, since any archive can use the checklist of criteria to measure its own performance and to identify any areas in need of improvement.

At the core of OAIS is a set of high-level functions which describe an archival system capable of preserving digital information and of making it available over the long term. OAIS also includes a set of responsibilities that an archive must fulfil. OAIS does not give detailed guidance about the procedures, protocols, and applications that are used in a preservation system; it is instead a general model, equally applicable to a national archive as to, for instance, a domestic collection of photographs.

2.1 Nomenclature

OAIS is designed to be applicable to a wide variety of disciplines, and so the model deliberately avoids terms which may have different meanings in different fields. This can make reading the OAIS recommendation confusing initially. In this document, OAIS terms have largely been avoided in favour of those more familiar to film archivists; however a few OAIS terms have been employed, in particular *Information*, defined thus:

Information The whole package of data and metadata for preservation (for instance, a digitised film, plus cataloguing and conservation data).

3.0 Mandatory Responsibilities

3.1 The Requirements

OAIS identifies the following mandatory responsibilities of a digital archive:

1. **Negotiate for and accept appropriate *information* from acquisition sources**
2. **Obtain sufficient control of the *information* in order to meet long-term preservation objectives**
3. **Determine the scope of the archive's user community**
4. **Ensure that the preserved *information* is independently understandable to the user community, in the sense that it can be understood by users without the assistance of the originator**
5. **Follow documented policies and procedures to ensure the *information* is preserved against all reasonable contingencies, and to enable dissemination of authenticated copies of the preserved *information* in its original form, or in a form traceable to the original**
6. **Make the preserved *information* available to the user community**

3.2 What they mean

Responsibility 1 means that the archive must reach agreement with the source of the material on a number of factors before accepting the material:

- What the *information* comprises (the extent of the collection, and what metadata – cataloguing information – is provided)
- Who is responsible for selecting and sifting the material, if required
- In what formats the *information* is to be supplied
- How the *information* is to be delivered

Responsibility 2 means that the rights and ownership must be established and agreed (and clearly recorded), and that these must not prejudice the archive's ability to preserve the material.

Responsibility 3 requires the archive to define who its principal users are, in order to be able to meet its obligations for collecting and disseminating its content. This might mean devising and maintaining an acquisition policy, for instance.

Responsibility 4 means that the archive must determine how much contextual information is needed so that the principal users are able to understand and use the material, and ensure that the users have access to this contextual information. Such information may include an explanation of how the material was created, and how it was intended to be used, as well as a description of its format and structure. To fulfil this responsibility, an archive must have sufficient resources to document material it acquires.

Responsibilities 5 and 6 oblige the archive to have clearly defined preservation objectives, to document the policies and procedures for carrying out these objectives, and to have a fully reliable and tested system in place for doing so. The content must be findable by the archive's users through suitable access mechanisms, and be

deliverable in a suitable format, whose relationship to the original form of the material is clearly defined.

3.3 Examples

- An archive contemplating the acquisition of a collection of digital sound and image files related to a production must first establish the range and content of the material, and whether the archive is authorised to sift and dispose of unwanted items. Also the archive must ensure that the formats are compatible with the archive's systems for preservation and access. The rights and ownership of the material must be clearly agreed before transfer, and must not prevent the archive from making copies for preservation purposes, and from disseminating the material to its users. If the material is in a proprietary format only usable in dedicated applications (such as Pro Tools audio files), then the archive must either convert these into formats more compatible with preservation and access, or retain sufficient information to allow interpretation of the material once those proprietary applications are no longer available.
- An archive acquiring digital cinema material must ensure that if a DCP is offered, that it is not encrypted, otherwise the archive will not have sufficient control of the content to preserve it, and cannot make it available to its users.
- Material in the archive which is of little interest to the archive's principal users might be considered for disposal.

4.0 The OAIS Functional Model

The OAIS functional model is rooted in the notion of *information packages*. An *information package* consists of the digital object that is the focus of preservation, along with metadata necessary to support its long-term preservation and/or access, bound into a single logical package. Note that this is not necessarily a single physical package, so that different elements of the package may be stored separately. OAIS defines three types of information package, the **Submission Information Package (SIP)**, the **Archival Information Package (AIP)**, and the **Dissemination Information Package (DIP)**. The SIP is the package received by an archive from the source. The AIP is the version which conforms to the archive's preservation standards and which results from the ingest process; this is the version stored and preserved by the OAIS, and may in some circumstances be identical to the SIP. The DIP is any version delivered to a user as the result of an access request. For a film archive, the SIP might be a DCP, a set of DPX files, an HDCAM tape, each with its associated metadata. The AIP is the preservation format (for instance DPX, JPEG2000), again with associated metadata (which includes preservation metadata), and the DIP will be whatever the archive's users require (such as a DCP, an H.264 video file, along with the metadata they need in order to use the item).

The functional model breaks down the elements of an archival system into the following components:

- **Ingest**
- **Archival storage**
- **Data management**

- **Administration**
- **Preservation planning**
- **Access**

4.1 Ingest

Ingest, in the broad sense used in OAIS, is the set of processes responsible for accepting SIPs into the system, and includes the means to receive the SIP, validation that the SIP is uncorrupted and complete, transformation of the submitted item into a form suitable for preservation within the OAIS (that is, the creation of the AIP), the extraction and creation of metadata (which importantly includes the generation of 'fixity' information such as a checksum which allows future protection against corruption), and transfer of the AIP to the archival store.

An archive should at the earliest stage of acquisition generate checksums for the digital material it receives. This is the only reliable means of complying with the fifth mandatory responsibility, namely ensuring that the material can be preserved in its original form, or in a form traceable to the original. Ideally this process should be carried out at the time the material is first received by the archive, before any other actions. This will ensure that at every subsequent stage the archive is able to check that nothing has been done to alter the integrity of the material received. Whether or not the archive does generate checksums as the very first stage, it is essential for long-term preservation that the AIP does include checksum data to allow future monitoring of the material.

Also, as part of ingest, the archive should validate that the received material is uncorrupted and complete. This may be by using verification software, or by manual checks, or by other means.

The archive should also gather all the available metadata, which may be embedded metadata or information supplied separately, and, if this is insufficient to allow management of the future preservation and access of the material, should create metadata sufficient for this purpose, including preservation metadata (see below).

A process must then exist to transfer all the gathered *information* (the AIP) to archival storage. Again, this may be fully or partly automated, or may be completely manual.

4.2 Archival Storage

This function is responsible for ensuring that archived content resides in appropriate forms of storage – eg. online, near-line, off-line – and that the bit streams comprising the preserved information remain unchanged over the long term. It also retrieves items from the storage system to support access requests.

The principles, if not the practice, of long-term archival storage are straightforward: the information must remain unaltered and 'renderable' (that is, readable) over the long term. This means not only that the systems used to retain the information are robust, that error-checking is periodically carried out, and that reliable disaster-

recovery procedures are in place, but also that formats are monitored and migrated whenever necessary to ensure that the content can still be accessed.

In practice an archive has to choose between a bewildering variety of hardware and software, and between automated and manual systems, in a world where digital formats are rarely ideal for long term preservation. The OAIS principles however offer the basic ground-rules: for instance, there is no bar to an archive choosing to store AIPs containing files a proprietary format, provided that these are continuously monitored and ultimately migrated to a new format before they become unusable (see Preservation Planning, below).

4.3 Data Management

This is the management of descriptive metadata to support the archive's finding aids and management of system performance and statistics. An archive must be able to manage and maintain databases which enable search and retrieval of the archive's content, as well as enabling the administration of the archive's internal operations.

4.4 Preservation Planning

Preservation planning is the monitoring of the changing environment and the updating of the preservation strategy in response. Preservation planning necessarily will include a strategy for migration of digital data from media and formats known to have obsolescence built in to their development roadmap (such as the LTO data tape format), even though the actual migration is part of the archival storage function.

4.5 Access

These are the processes by which users locate, request and receive items from the system. Such processes include processing database searches through the Data Management function, coordinating the delivery of the requested content from Archive Storage, and processing the material into a form suitable for the user (which might include transcoding and stripping away inappropriate metadata). An archive's principle users must be able to search the system, and gain access to the content in a usable form, without the need for assistance from the original source of the material.

4.6 Administration

Administration is responsible for managing the day-to-day operations of the archival system, as well as coordinating the activities of the other five high-level functions.

5.0 Preservation Metadata

Preservation metadata is the information an archive uses to support the digital preservation process, and is an essential part of the AIP. Preservation metadata includes information about the digital item, about its creation or modification, about rights, and about the people, organisations or applications involved.

A number of organisations worldwide support a standard for preservation metadata called *PREMIS*. Although there is no absolute requirement that a digital archive complies with this, PREMIS does offer the *PREMIS Data Dictionary*, a formal framework which can be used as a reference in the implementation or evaluation of a preservation system. To comply with PREMIS, an archive must be able to export data

from its system which conforms to the Data Dictionary, even if the data is held in another form within the archive's own system.

6.0 Trustworthy Repositories

ISO 16363: Space data and information transfer systems – Audit and certification of trustworthy digital repositories provides a checklist as a means of measuring the performance of a digital archive against a standard set of criteria based on OAIS. It is invaluable as a means of assessing an archive's reliability and of identifying weaknesses in its approach to digital preservation.

The audit criteria are divided into three sections:

A. Organisational infrastructure
B. Digital object management
C. Technologies, technical infrastructure, and security.

Within each of these sections are the criteria an archive must comply with in order to be trustworthy. Although these are expressed using OAIS nomenclature, each one is carefully explained with examples, making this a powerful tool which can be used as part of the process of setting up a digital archive, or for assessing the trustworthiness of an established one.