Where the OAIS Ends: Archival Principles and the Digital Repository

Maygene Daniels; Gallery Archives; National Gallery of Art; Washington DC/USA

Abstract

The OAIS reference model (ISO 14721:2012) has been widely accepted as a theoretical foundation for digital archives, but there has been relatively little discussion of internal elements of an OAIS-compliant repository in the context of traditional archival theory and practice. Based on the experience of the archives of the National Gallery of Art, this paper presents practical concepts for creating a small digital archival repository within the OAIS framework built on recognized archival principles of provenance, group level management, and hierarchical organization.

The OAIS and Archives

The Open Archives Information System - the OAIS [1] - has been widely accepted in the archival community as a compelling theoretical framework for a successful digital archives. Its requirements for managing digital data based on institutional responsibility and service to defined user communities closely parallel traditional archival practice, and the OAIS functional model tracks core archival tasks of appraisal and accessioning, processing and preservation, and reference service, albeit with different terminology.

The OAIS information model, which conceptualizes a preservation unit or "package" as one that integrally links the digital object with metadata also is easily understood in the context of traditional archival practice. The OAIS further defines three types of "packages", the *Submission Information Package*, or SIP, the *Archival Information Package* or AIP, and the *Dissemination Information Package* or DIP, all of which also are mirrored in facets of modern archives administration.

For all its clarity, however, the OAIS is a theoretical model that intentionally avoids any statement on how to go about creating a compliant digital archival repository. Although most institutions already face an immediate need to preserve digital data, few practical models are available. While many experiments are underway, best practices are only now emerging.

My goal in this paper is to draw on the experience of the Gallery Archives of the National Gallery of Art to suggest an approach to creating a low-cost digital archival repository within a small institution [2]. The OAIS provides one context for the repository, but traditional archival principles are equally important guides. These familiar concepts require that records that belong to one records creator should not be combined with documents from another source and that the context and relationship of individual documents within groups of associated materials should be respected and guarded [3].

Although bit-stream security is essential for a digital archives, this responsibility is ably managed by the Gallery's IT department and so is not discussed extensively here. Instead, this paper focuses on internal structures and strategies that can be used to create a reliable digital repository within the OAIS framework and a traditional archival context.

Repository Characteristics

Some of the characteristics of a successful digital archival repository are clear.

Following the OAIS model, external metadata must be complete and integrally linked to digital objects. To ensure long-term preservation and access, proprietary and undocumented software should be avoided. To promote the greatest likely longevity, file formats should be limited to widely used open formats. To ease future format migration, file formats should be limited and standardized.

Of equal importance, the structure of the repository should be robust and flexible. It should be consistently and logically organized and digital names should be humanly readable so that digital objects can be understood and located without intervening software. Excellent descriptive metadata – finding aids in the archival context – is also critical and an expected part of any archival repository.

Evidently careful security is essential for a successful repository. To help minimize human error, access levels should be defined and monitored. Policies should require oversight during critical operations. Should an error occur, distinctive naming conventions for folders and digital objects and redundant information throughout the repository make it easier to correct the problem.

Other factors also have become clear over time. One important realization is that a successful digital repository should be as simple as possible. Funding and staff are always limited, heavy institutional responsibilities already exist, and digital preservation presents a major new initiative. For economy and consistency, existing archival management systems should be used wherever possible. Similarly, free and inexpensive small-scale digital tools can be used to move and rename files, and simple scripts can effectively automate large parts of the process. Although a major new initiative may be impossible in times of financial stress, incremental development is generally achievable. Small steps succeed where the grand may fail.

Perhaps most important, a digital repository must be understood as an integral part of the larger archival repository to which it belongs. Even as record-keeping practices evolve inevitably toward digital media, archives continue to be responsible for legacy analog formats and for digital versions of analog originals. A successful digital repository must relate seamlessly to the entirety of the institution, and the relationships between materials in multiple formats must be self-evident.

Repository Structure and Workflow

Responding to these concerns, the Gallery Archives has built a simple low-cost digital archival repository within the OAIS framework. The repository is lodged in a data share on one of the Gallery's servers [4]. At the highest level, the data share and the digital archival repository are divided into six functional sectors, actually high-level network folders. The first sector is for ingest, the second is for temporary storage and processing, the third and fourth sectors are for aspects of permanent preservation and access, the fifth serves as a records center for less valuable materials, and the sixth sector is flexibly used for temporary storage of reference materials.

Figure 1 - Digital storage sectors

Sector 1 - Ingest
Sector 2 - Working Space
Sector 3 - Digital Archives
Sector 4 - Preservation Archives
Sector 5 - Records Center
Sector 6 - Reference Storage

Let me describe the workflow in the digital archival repository.

Ingest begins in Sector 1 when, following initial planning and discussion, creators deposit digital objects into a shared network transfer folder with the original digital names and folder structure intact [5]. This constitutes the *Submission Information Packet* or SIP in the OAIS context. The department notifies the archives that the digital materials are available for pick-up and provides supporting information about the transfer, including spreadsheets and descriptive data. Archival staff examine the digital files to make certain that they have been correctly and completely delivered and are uncorrupted, with adequate technical metadata.

If everything is in order, the archives assumes physical possession and ownership of the files and, following procedures that are used for all acquisitions, creates a record of the transfer in the archives relational database. From this point, the acquisition is identified by the unique transfer ID assigned by the database.

With the transfer number as identifier, the digital files are then moved intact from Sector 1 to Sector 2 - the Archives Working Space - where they can be held, evaluated, and processed [6]. Depending on the archival workload and competing priorities, the digital assets can be securely maintained in the working space as long as necessary.

The next steps in the digital ingest process are human and intellectual. Processing the digital materials begins with a deep and thorough evaluation of the nature and long-term significance of the files. The content and structure of the digital objects, their relationship to other records, their usability, and the value of information or evidence they contain all are evaluated to determine whether the digital files should be preserved permanently. In many respects this judgment is comparable to traditional archival appraisal, although with greater focus on technological form and long-term usability. If a group of files lacks sufficient value, it is moved to the records center sector of the archives digital data share. The digital objects are documented and maintained there with the assigned transfer ID for tracking, but files are not renamed nor is detailed metadata created.

Other files, those determined to have long-term value, are prepared for archival storage. In OAIS-terms, they are converted from a *Submission Information Package* or SIP into an *Archival Information Package* or AIP. This process closely parallels steps in archival processing of physical materials.

To summarize briefly, each new acquisition is assigned to a record group and each group of functionally and physically related materials within the transfer is assigned to an archival series.

The existing digital folder structure is analyzed and a spreadsheet is created with original folder names and revised folder titles developed by archivists in accordance with archival conventions. The spreadsheet is imported into the archives database to create new folder-level records each with a unique system-assigned ID. These folders are linked to the appropriate record group and series so that all files and items that were created and used together and share a physical form are associated with one another.

The next stage of processing concerns the items at the hierarchical level below the folder. Digital object names are harvested on a spreadsheet and further description is added if needed. The spreadsheet is imported into the database to create item records. If possible, original item names are assigned to the individual digital objects, but if none are available serial numbers are applied.

After folder and item records have been created, the archives database becomes the source of the universal repository-assigned digital file names. These are perhaps the most important element of the digital repository. The system can be extended to any item in the archives and works equally well for digital or analog originals and digital copies of analog materials.

The complex names consist of the record group and series names, a database-assigned folder ID, and the item designation, separated by underscores. An example of an identifier for a digital object would be 26B7_3303_C79-17-X-8.tif. 26B7 indicates that the item is a digital scan of an exhibition installation view, 3303 is the folder ID linked to extensive metadata, and C79-17-X-8 is the original catalog number assigned by the office. The original object from which the scan was made would have the series designation 26B6 (exhibition installation slides) but would otherwise be the same except for the final extension.

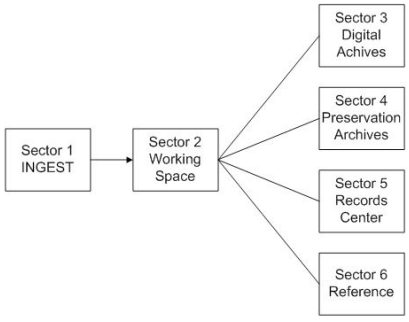

Figure 2 - Functional diagram of digital archival storage

One further wrinkle of the digital archival repository deserves further explanation. Although often digital materials are

transferred to the archives in a single format, they may be received in multiple formats all of which merit preservation. As an example, a photographer might transfer images both as camera raw files and in processed TIF form. Similarly digital scans of textual materials can be delivered both as searchable PDFs and native TIF images. In each instance one of the versions, in the first case the TIF version and in the second case the PDF, is intended to provide convenient access while the second format is needed for preservation but not for regular use. Recognizing this distinction, as a point of convenience and efficiency, the digital archival repository includes a special sector called the preservation archives to hold secondary forms where they exist.

In the final stage of the ingest process, archivists apply scripts and use external cataloging tools to rename and move digital folders and objects with long-term value to the appropriate digital preservation sector. The renamed folders self-sort by series and folder hierarchy, and related items naturally fall in correct order. Thus in many ways the digital archival repository becomes a close analog to a traditional archives.

From this point, the digital assets are flexibly available for access and viewing using software outside the digital repository. At present the archives uses a digital asset manager to view visual items and Adobe Bridge to read textual documents. Better viewing systems are expected in the future. In keeping with the OAIS concept of the *Dissemination Information Package* or DIP, the format or size of digital objects may be changed for delivery to users but key descriptive metadata is maintained so that the meaning and context of the digital objects are always clear.

Conclusion

Evidently this brief discussion can only give the general outlines of a still-new digital archives [7]. Following the OAIS model, the repository provides firm, reliable control of digital assets, and makes them readily available to non-invasive access tools. Its systems are infinitely scalable, and its organizational structure and naming conventions adhere to archival principles. The repository has been successfully expanded to include a wide range of physical materials and equally has accommodated a geometric growth rate. While now located on a shared server, parts of the repository could easily be assigned to near-line or off-line storage if needed.

Yet significant challenges remain. The blending of intellectual and automated processes remains complex, and greater automation will be needed to manage the ever-increasing volume of digital transfers. Lying ahead, direct relationships with electronic records-management systems must be established and the repository will need to manage more complex digital objects and an ever-larger volume of media.

Despite future challenges though, at present the repository succeeds in its principle task. Digital materials are being preserved securely now so that they will be available in the future as new systems and best practices evolve.

Notes

[1] The OAIS became an ISO standard in 2002 and recently has been revised and published as ISO standard 14721:2012. "The Open Archival Information System Reference Model: Introductory Guide" by Brian F. Lavoie (Digital Preservation Coalition Technology Watch Series Report 04-01, 2004) is an excellent plain-language guide to OAIS complexities.

[2] The National Gallery's digital archival repository initially was created to manage visual objects. As requirements have evolved, the repository has expanded successfully to manage digital copies of analog and increasingly born-digital textual and graphic documents and media

[3]Creative development of the repository has been a collaborative effort of many individuals in the Gallery Archives and other departments at the National Gallery of Art. Greg Swift, Tom Walton and Chuck Patch provided critical insight on server organization and naming conventions from an IT perspective. Christina Waldron gave expert guidance on database management and authored scripts for key points in the ingest process. In the Gallery Archives, every staff member contributed insight and practical solutions as the digital repository developed. Michele Willens was the primary project manager. Julie Blake helped visualize and document the ingest process.

[4] The server is RAID-protected and backed up nightly, with a regular schedule of off-line, off-site security retention managed by the IT department.

[5] If for some reason network transfer is impractical, the archives also accepts removable media, most commonly DVDs, which archives staff then upload to the server.

[6] For security, a redundant copy of the folder and its contents also is created on a high capacity off-line hard drive so that at any point in the process, the original asset can be recovered. Any original transfer media also are preserved as deep backup.

[7] The Gallery Archives received its first digital transfer via floppy disc in 2004, a project to create digital preservation scans of analog originals got under way in 2008, and for the first time this year discussions are beginning for transfer of large amounts of born digital graphic and textual documents.

Author Biography

Maygene Daniels is chief of the Gallery Archives of the National Gallery of Art. She has served as president of the Society of American Archivists, president of the Academy of Certified Archivists, and chairman of the International Council on Archives Section on Architectural Records. She is a fellow of the Society of American Archivists.

