

INSTITUTIONAL READINESS

Every library arguably owns material worth digitizing and serves audiences likely to have growing appetites for electronic information. In this environment, organizations are understandably eager to undertake digitization projects.

Learn-by-doing, in-house strategies are attractive. Librarians are often tempted to purchase a scanner, select a collection to digitize, and move forward. Development of websites and exhibits, digital document delivery via interlibrary loan, article delivery in e-reserves: all demonstrate the ease of digitizing library materials.

Begin digitizing *collections* too soon, however, and projects can fail by succeeding.

Succeeding in making digital copies is one thing, but identifying, storing, retrieving, and delivering digital objects within a library context—both as independent entities and as parts of a collection—is another. Digital files tend to be stranded on disks—and potentially forgotten—when scanning occurs well before digital collection management and delivery services are established.

From the user's perspective, digital surrogates have no value unless they can be reliably located, retrieved, and delivered with appropriate information for identification and interpretation.

Capabilities to catalog (identify), store (preserve), deliver (circulate), and interpret (public service) are traditional hallmarks of libraries. If the purpose of digitization is to create sustainable library collections—or to contribute items to existing digital collections—then an organization should first assess its capabilities to provide post-acquisition library services for electronic formats before embarking on creating them.

Assessing an institution's policies and procedures for acquiring and delivering digital materials is a useful measure of an organization's readiness to create digital resources. Consider the following:

- Are there established conventions for cataloging and intellectual control?
- What procedures are followed to ensure that the library has the right to distribute a given digital resource?
- *Is there a place to put the material?* And a database the collection manager can use to aggregate and retrieve objects or their component parts?
- Once acquired, cataloged, and stored, can the resource be named? In a persistent fashion?
- Are delivery applications in place to render (that is, display) objects and metadata properly?
- Are tools available to assist the user in studying, navigating, interpreting, printing, or making other uses of the displayed object?

In sum, should librarians collect digital materials, even at a modest scale, without yet having the library in which to place them?

The same criterion applies to producing digital collections. Digitization programs have the best chance for initial success when they are established at the right time.

Build the library before the collections

The range of viable technologies to store and deliver digital reproductions is both liberating and vexing. Books, for example, can be converted to digital images (page images) in black-and-white or color; delivered with or without searchable text; and presented with or without capabilities to go to specific page numbers or click from section-to-section.

The most efficient workflows follow specifications that convert collections once. Among all digital product choices that are available, the best are the ones that fit the needs of the library's audience and the library's infrastructure.

Catalog records, structural metadata, text files for searching, and digital images for display all should be produced in formats known to the database, storage, and delivery applications that the institution has invested in or is about to purchase.

Beware of choosing formats that prove to be incompatible with downstream applications to manage and deliver catalog records and digital reproductions. More than one library has encountered the need to reformat (again) just-digitized material to meet the goal of delivering functional objects to users.

During the past 10 years, libraries have adopted two strategies to institute digitization. Either a library plans a pilot project designed to create content *and* infrastructure, or the initial project is content-driven, with the expectation that infrastructure will follow (and be separately funded).

The Andrew W. Mellon Journal Storage project (JSTOR) is a noteworthy example of an initiative to build content and infrastructure. Through two phases of pilot projects and appropriate levels of investment to create large-scale infrastructures for storage and delivery, JSTOR succeeded in meeting deadlines to digitize and distribute back runs of journals.

Note that initial investments for infrastructure in this project were much greater than ones for services to digitize the journals and deposit the files into the newly created JSTOR system.

Content-driven projects, on the other hand, yield digital reproductions, but not necessarily the means to identify and distribute them.

Of the many projects undertaken in the 1990s to convert photographs and other visual materials to the Kodak PhotoCD format, how many of these images were made and continue to be available via the Internet? Any library office that has one or more Photo CD disks on the shelf today exemplifies a successful low-cost approach to digitization but not necessarily digital library collection development.

Downstream technical infrastructure

Following scanning and cataloging or production of other descriptive metadata, the technical infrastructure needed to inspect, edit, manage, sustain, and deliver digital collections is extensive:

- Networks and file servers with sufficient capacity to move digital assets to multiple locations (for quality control, assembly of multipart objects, or other tasks) during production

JSTOR, www.jstor.org

- Public catalogs that facilitate searching for digital surrogates in their appropriate context (Would a library digitize personal papers or other manuscripts described in a finding aid without having the means to index and search encoded finding aids? Or digitize photographs without yet having a database or online catalog that permits item-level searching of images?)
- Robust storage systems and databases to store completed digital objects and their associated administrative and technical metadata
- Authorization and authentication systems for content restricted to domains smaller than the entire world
- Naming and name-resolution systems for persistent linking
- Web servers and delivery applications (user interfaces)

Fortunately, this full infrastructure does not need to be replicated in every institution. The Colorado Digitization Program, for example, centralizes storage and maintenance of collections digitized from institutions statewide.

Libraries that have contributed content to the Library of Congress American Memory collections demonstrate the viability of relying on the technical infrastructure of a partner organization—either as sole or adjunct provider of management and delivery services.

Policies for ownership, maintenance, and distribution

Policies for ownership and physical control of surrogates largely influence the geography of digitization projects. Policies regarding control of the bits help answer one of the essential questions of workflow: Where will the data go?

Organizations without repositories or repository-like infrastructures (comprised of online storage systems with error checking, reliable backup, and associated databases to store administrative metadata) might find purchasing storage from a service provider such as the OCLC Digital Archive or from a partner in a multi-institutional initiative advantageous.

By ceding physical control over some or all its digital assets, a library can delegate storage responsibilities (bit preservation) to another entity. Any cost savings realized by purchasing instead of building managed storage services could then be used to fund other program components, such as digitization, cataloging, or delivery.

Establish ownership and maintenance responsibilities before digitizing because entities that offer back-end storage might prescribe or recommend best formats to deposit data objects. They also may require the owner to provide technical metadata that could more easily be captured during production than afterward. If known before digitization, these repository specifications can be incorporated into the digitization specification and workflow.

Providers of delivery services are even more likely to impose strict terms and conditions on contributing material to their databases. Image consortia such as Amico and ARTstor specify in detail the formats for metadata and digital images to be deposited.

When weighing the options of using external services to store digital masters, or to host delivery versions available to users, preservation responsibilities must be understood and documented. If any of the owning library's responsibilities necessitate obtaining copies of the data—even if only for brief periods—the question of Where the data will go? will be raised again.

Amico, www.amico.org

ARTstor, www.artstor.org

Ingrain concept of sustainability

Digital objects are fragile. Preservation requires advocacy. The best time to build in sustainability is at the point of creation.

Rates of deterioration are notoriously difficult to demonstrate, and even more difficult to predict for electronic information because of the many variables that contribute to obsolescence.

Plans for programs designed to sustain digital collections should implicitly, if not explicitly, acknowledge three factors:

- Regardless of the methods used, *digitizing* is not synonymous with *preserving*. Short of solving one problem (for example, replacing brittle paper), digitization creates another. Whenever source materials are retained, producing surrogates at least doubles collection size—or triples it when workflows yield master and delivery versions—and significantly increases the amount of *active* oversight required to monitor obsolescence.
- Media longevity cannot be measured independently of the storage environment. Where the collections are stored is as basic to longevity as the quality to which they are produced. The American National Standards Institute (ANSI) subcommittee IT9-5 has recently issued standards-based methods to measure life expectancy for compact discs (ANSI/NAPM IT9.21) and storage recommendations for magnetic tape (ANSI/NAPM IT9.23). Both underscore the strong relationship between media and environment when projecting life expectancies.
- Sustaining use constitutes a larger preservation mandate than sustaining content.

Discussing sustainability at the outset of program development—before undertaking digitization projects—reveals individual biases and assumptions within the organization. Do not assume that all stakeholders in digitization initiatives will understand or agree to a single definition of sustainability.

When addressing the distinction between *sustaining use* versus *sustaining content*, account for the changing expectations of users in addition to technology changes as a key challenge to sustainability.

Microfilm is a case in point. Any of the following factors *could* be cited as signs of technological obsolescence and reasons to be concerned about sustaining operations to produce microfilm:

- The decreasing availability of film stock
- Decreasing film processing services
- Difficulties of maintenance for legacy equipment
- Discontinuation of reader/printer product lines

Assuming, however, that a library has legacy film, is responsibly storing film masters (in cool storage), and is maintaining its service infrastructure for service copies of film, how would the organization assess fulfillment of its (stated or implied) preservation obligation if users abandoned the format?

What happens when the media are there—readily identified, well stored, supported by delivery systems—but no one comes to use them?

Precedent exists for users to reject one stable medium in favor of newer higher-quality or more convenient formats. Consider the fate of lantern slides when 35mm slide film and projectors became widely available. With the imminent disappearance of slide projectors, 35mm slides will soon be technically obsolete (although physically preserved).

An organization's infrastructure and policies should accurately reflect its (local) concepts of sustainability. When the obligation is to preserve the artifact (sustain content), invest in stable media and controlled environments.

When the obligation is to sustain use, however, then presumptions about longevity, as well as post-digitization budgets, need to acknowledge the owner's obligation to make additional investments—for example, to reformat material—well before the physical life expires. Administrators should acknowledge that intervals between interventions are likely to be shorter in the “sustain use” versus the “preserve the artifact scenario.”

Whether an organization seeks only to sustain the files it produces in digitization (preserve the bits and preserve the artifact) or to sustain use (preserve the work), the sensible digitization strategy is to create digital collections in optimal formats placed in good environments.

Administrators should strongly advocate the per-item, long-term cost effectiveness of optimal storage environments—where *optimal* could be measured, for example, against the Open Archival Information System (OAIS) Reference Model standard's entities for “Data Management” and “Archival Storage.”

They also should caution stakeholders that even under ideal storage conditions, collections governed by a commitment to preserve use will require transformations at intervals that keep pace with changes in technology or user expectations.

In practice, the organization's obligation for sustainability, its policies for ownership and control of the data, and its choice of service provider for storage dictate which storage media, object, and technical metadata formats are best to meet program needs.

Develop funding models

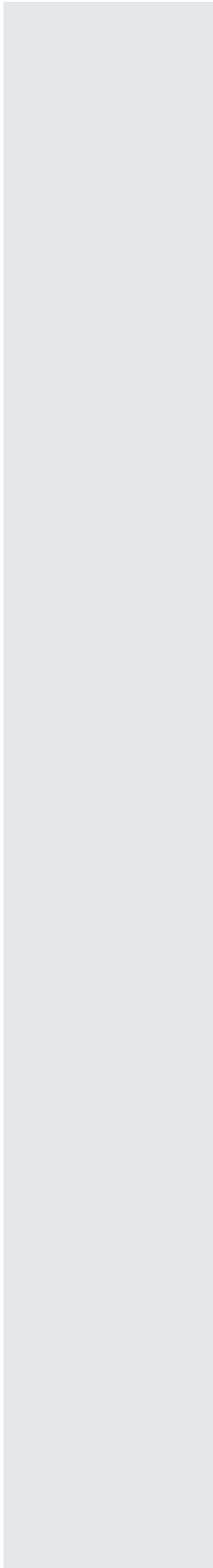
A June 2000 survey of Association of Research Library (ARL) members reported that half the funding for digitization came from operating funds, half from grants.

Recent publications from OCLC and the Council of Library Information Resources (CLIR) provide economic models and business plans respectively as means to sustain digital assets.

Given the high costs to create and, at present, to sustain digital collections, the pressures are great for cultural heritage organizations to design and implement viable funding models. Collaborative approaches to digital presentation may prove to be an arena, like the development and maintenance of union catalogs, in which libraries participate in for- or not-for-profit consortia to distribute costs.

In the meantime, as a matter of high-level planning to create a digitization program, the administrator should recognize that stable funding is just as important as a stable technology infrastructure to creating good digital library collections.

Survey, “Preservation and Digitization in ARL Libraries,” July 2001, www.arl.org/spec/262sum.html



To lobby for *some* of the funding lines—for staffing, technology, purchased services—needed to put a program onto stable footing, the administrator might argue that in the increasingly competitive environment for external funding, an institution with infrastructure will be in a much stronger position to obtain grant funds for digitization projects.