

by
marshall
breeding



The Systems Librarian

► Independent Consultant and Founder of Library Technology Guides

Ongoing Challenges in Digitization

THE ENDURANCE
OF SCHOLARLY
AND CULTURALLY
IMPORTANT CONTENT
DEPENDS ON
LIBRARIES AND OTHER
ORGANIZATIONS
MAKING LONG-TERM
COMMITMENTS
TO DIGITAL
PRESERVATION.

In the course of my career, I have been fortunate to be involved in a variety of digitization projects. During my tenure at the Vanderbilt University's libraries, I led the digitization of its Television News Archive collection of 50,000 hours of video and participated in many other projects involving digital audio (such as the Global Music Archive, archival documents for the Ecclesiastical and Secular Sources for Slave Societies, as well as many photographic collections). In my travels, I've had opportunities to observe and learn about the digitization operations in many national libraries and other cultural institutions around the world. My interest in digitization is carried out in my personal life as well, through the scanning of thousands of photographs taken by family members through the years and managing an even larger number accumulated since the advent of digital photography.

It has been interesting to see the tools and technologies evolve through the decades. The costs of storage and equipment have dropped dramatically, and there are more applications and platforms for managing digitized collections, ranging from those available as open source to more expensive high-end products. The quantities of digital content are exploding, which presents many challenges for storage, management, access, and preservation. Through the digital projects carried out in libraries, the body of interesting and important content available to scholars and the general public has seen dramatic expansion.

Digitization and Preservation

One of the key roles of libraries and cultural institutions is preserving resources for future generations. Today, we benefit from manu-

scripts, books, and other artifacts that have survived for hundreds or even thousands of years. Given the right safeguards, environmental conditions, and the absence of catastrophic events, physical objects can be made to endure for long periods of time.

Digital objects require considerable effort to ensure their long-term preservation. Any storage medium used for digital files may have a life expectancy measured in years or possibly decades, but certainly not centuries. Digital files are fragile and require constant attention if they are to endure in the long term. Making digital content available to scholars and researchers who might live a few hundred years in the future relies on a constant and unbroken chain of migrations to new storage media, file formats, and controls to ensure and correct the integrity of all the bits that comprise each digital object. Scrolls hidden in a desert cave that have survived hundreds of years are available to those able to read ancient languages. I'm confident that any digital media deposited in even the best environmental conditions would not be accessible for that same interval. The media would have deteriorated, the files would be corrupted, and even if they were to be fortunate to survive intact, it seems unlikely that computer equipment in the distant future would be able to access obsolete technical formats. Endurance of digital content requires not static storage, but an ongoing process of forward migrations.

Given the fragility of digital content and the risks inherent in long-term digital access, the best path for preservation should have physical and digital alternatives whenever possible. I can imagine very few circumstances in which a unique physical object must be discarded once digitized. While both physical and digital paths

have their risks and challenges, parallel efforts provide the most security.

There are cases, however, in which the options for preserving physical representations of content may not be viable. For the previously mentioned Television News Archive, for example, we faced the need to digitize more than 50,000 hours of content that was in jeopardy. The collection resided on a type of analog tape that was obsolete. There were a diminishing number of playback machines in existence, and we were approaching the life expectancy of the tape medium. Without intervention, an incredibly valuable collection would eventually become inaccessible. We set out to convert the collection into a current digital format that would both provide the basis for long-term preservation and flexible options for access. We kept the original videotapes, but the time will come when they will no longer remain viable points of access to that content, leaving the full burden on the digital versions.

Another scenario involves print manuscripts in such jeopardy that digitization may be the only practical route for their survival. I was involved in a

project to digitize volumes from archives in Latin America where the environmental conditions in which they were stored were abysmal with no real prospects for providing better conditions in the future. Many of the volumes were already water-damaged and worm-eaten. Digitizing provided the best way to capture the content before it deteriorated further. The project participants—often historians or graduate students—captured each page with handheld digital cameras, since the authorities of the archives allowed only a brief period of access and the work had to be done on-site. These images have proven to be a valuable resource for the historians researching the migration of slaves from Africa through Latin America and to North America in the 18th century.

For digitally native content, digital preservation stands as the only alternative. We live in a time in which most new content is created digitally. In some cases, there might be the possibility for creating physical representations for preservation. However, in practical terms, these materials will depend on digital preservation processes to survive. The endurance of scholarly and

culturally important content depends on libraries and other organizations making long-term commitments to digital preservation.

Funding Digitization Projects

Despite what I mentioned previously about the decreasing costs of storage and equipment, digitization projects require considerable resources to execute. Even if infrastructure costs diminish, these projects require considerable staff time. Anyone who has been involved in one of these projects knows that the time and expertise involved in creating metadata, transcriptions, or other descriptive content far exceeds that expended for the digitization itself.


An increasing number of libraries have digitization departments as part of their routine operations. It's no longer considered exceptional or extraordinary for libraries to be digitizing items out of their special collections, but it is more of an expected activity for those institutions that have substantial unique materials. The mass digitization of books is another matter, which does generally involve external funding and coordination.

Write for Us in 2015!



Here's what we've been talking about lately in *CIL* magazine. But don't let it limit you.

We're always looking for case studies, firsthand accounts, research reports, and tips-and-tricks articles about how librarians have applied technology in order to better serve their communities, colleges, countries, and companies. Submit your story ideas anytime at infotoday.com/cilmag.

 Information Today, Inc. 143 Old Marlton Pike, Medford, NJ 08055

In the past, digitizing collections was seen as something that a library might pursue only if it was able to obtain special funding. Many of the early digitization projects in libraries were carried out through grant-funded projects. Many of these projects aimed to not just result in the digitization of a given collection, but to build capacity in terms of equipment and expertise for ongoing work beyond the term of the grant. We are well past the time when digitizing a collection can be considered an especially noteworthy activity that might attract the attention of grant organizations. Grants may still be available for exceptional collections that might gain broader impact or that require digitization to ensure their survival as well as resources beyond those the responsible organization can muster.

Libraries seeking external funding for digitizing will have to work harder than ever to present a convincing argument. One angle for seeking funding could be designing a project that not only processes a given collection, but advances the art of digitizing in some way. Projects that are more likely to gain favorable attention from grant funders are those involving the creation of new open source tools, exploring new technologies or techniques for digitizing or describing challenging collections, or developing other innovations.

Push the Limits of Exposure and Access

Digitizing a collection provides many more opportunities for access beyond what is possible with physical artifacts. Digitization makes it possible to enable global access to everyone who might have an interest in and could benefit from being able to view and study those materials. Although digitizing materials enables the potential for global distribution, legal or practical barriers may stand in the way. Copyright concerns may prohibit access beyond in-library computers, for example.

Even when the full representation of an item may need to be restricted, it may be possible to provide access to some representation that will enhance its exposure of the collection and of the

library itself. Providing access to metadata describing the materials can be thought of as the baseline of exposure, which should be allowed for almost any item of content. There may be rare instances in which an object is classified or so restricted that even its existence cannot be exposed. But for the vast majority of materials, providing access to the metadata through as many channels as possible informs the library's own community about the materials as well as informing those further afield.

Implementing basic SEO techniques or semantic web technologies—such as that of Schema.org discussed in last month's column—will boost the exposure to the collection. How these techniques are implemented may vary depending on the platform used to manage the collection. Some of the basic SEO practices include providing a unique and persistent URL for each digital object, providing comprehensive itemization of the objects available in an XML file that conforms to the specification at sitemap.org, and exposing structured metadata in both the header of the webpage presenting the metadata as well as in the body of the page through microdata structures, such as those defined by Schema.org. In my experience through many different projects, attention to details that increases the exposure of a collection through web search engines results in increased use and potential value to the organization.

Beyond exposing metadata, it seems beneficial to provide access to the best possible representation of the digitized object. When digitizing a photograph, for example, multiple variants will usually be produced in the digitization process. These range from a minimal thumbnail image, to lightweight versions suitable for viewing through a web browser, and to ultra high-resolution preservation quality digital files. While libraries might consult with their own intellectual property attorney or advisor, in most cases, the small thumbnail of an image can be considered as metadata and therefore is available for free access. Whether full-sized access versions or preservation-quality mas-

ter files can be made available will be subject to decisions relative to copyright enforcement and other policies.

When copyright issues legally allow it, an important question is whether the library should provide access to its highest-quality images. Many libraries, special collections, or archives choose not to release the highest-quality images because they have revenue-generating licensing programs, or they are concerned that the images would be used or altered in ways they might not want to approve or condone.

My experience has led me to believe that providing access to the highest quality digital version will enhance its revenue-generating possibilities rather than diminish them. As long as any digital object includes a clear statement about the Creative Commons license that corresponds to the use allowed, most publications or other projects able to license the content will be more likely to do so if they can preview usable images.

Challenges Remain

Creating or providing access to digital content represents one of the most interesting and important activities in which libraries engage. I appreciate the part that libraries play in continually expanding the digital content available within their own communities and at the global level. On one level, these projects can be considered mainstream and routine. And yet, in order to build and sustain these collections, the tools and technologies need to continue to advance. I am especially concerned about the vulnerability and the fragility of digital objects. I hope that libraries and other organizations will invest in the preservation infrastructure and processes to ensure today's digital heritage remains available to distant future generations. ■

Marshall Breeding is an independent consultant, writer, and frequent library conference speaker and is the founder of Library Technology Guides (librarytechnology.org). His email address is marshall.breeding@librarytechnology.org.

Copyright of Computers in Libraries is the property of Information Today Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.