Enduring Access to Special Collections: Challenges and Opportunities for Large-Scale Digitization Initiatives

As we explore the evolving information landscape and institutional context of rare and manuscript collections, one of the critical matters is to consider the implications of large-scale digitization initiatives (LSDIs) for our programs. Although most LSDI efforts thus far have focused on general collections, it is inevitable that the attention will soon be turned to special collections. With the current networked information environment and increasing reliance on digital content subscriptions, rare and manuscript collections increasingly define the uniqueness and character of individual research libraries. The goal of this article is to characterize current LSDIs and discuss the potential implications for special and rare collections based on the insights gained through the first phase of LSDIs that have entailed general collections.

The digitization of millions of books under programs such as Google Book Search and Open Content Alliance has dramatically expanded our ability to search and find scholarly materials, especially books and serials. Research libraries and archives have been profoundly affected by the increasing availability of scholarly information on the Web, as this trend sets user expectations and behavior in discovering and accessing collections. To date, Google has defined the LSDI programs due to the scale of its operation as well as the publicity it has received from the program's ambitious and controversial nature. Currently representing 12 million digital books, Google continues to define the key parameters in converting published scholarly content into a searchable digital corpus.

The key value proposition of LSDIs for libraries and archives is the ability to capitalize on the changing means by which users discover, search, and use collections. Academic professionals recognize that discovery happens at the search engine level, as these are the tools most students and researchers use to start their research. It is only through brand-name recognition that users go directly to information portals such as JSTOR or Project Muse without the mediation of a search engine. Therefore, digitizing institutional assets and making them available on the Web so they

can be discoverable by search engines is seen as prerequisite to broadening access, and it is feared that scholarly materials not thus identifiable or accessible will be underused and neglected.

An additional incentive to making institutional collections visible worldwide via LSDIs is the transformative effect access to digital books has for several disciplinary and interdisciplinary studies. For instance, the aggregation in a virtual space of disparate and geographically distributed collections allows scholars to study social and national identities without being bound by nationality, race, ethnicity, or class. Research domains such as transnationalism now focus attention on movements and connections among people and ideas that transcend national boundaries and specific historical periods.

Another impetus to large-scale digitization is recognition of the Long Tail principle discussed by Chris Anderson in his 2004 *Wired* article. There is an assumption that, when freed from the physical confinement of bookshelves, every book will find a user in a networked environment. Research libraries and archives contain deep and rich collections. Even with digital access, however, users can be hampered in locating and obtaining materials of interest because institutions use different library management systems, with varying discovery and retrieval mechanisms. The continuing reliance on local library information systems results in confined metadata environments and requires users to interact with multiple interfaces. Anderson argues that the 80/20 rule (that is, that 80 percent of outcomes come from 20 percent of input) exists in the physical world because we chop off the "long tail"; in other words, the physical inaccessibility of an out-of-print or obscure work limits the demand for it.[1] On the basis of this argument, Lorcan Dempsey makes a compelling case for aggregating supply and demand at the network level rather than at the level of individual libraries.[2] An analysis of circulation records for materials chosen for Cornell University Library's Microsoft initiative showed that 78 to 90 percent of those items had not circulated in the last 17 years. In Cornell's case, the circulation frequency may be lower than average because of the age of the materials sampled: all were published before 1923. Nevertheless, the findings support the general perception that many of the materials in the library stacks are seldom used. This principle applies even further to rare and special collections due to provisions that prohibit the borrowing of such valuable and rare materials.

Except for a few initiatives, such as the ones carried out by the Smithsonian Institute and the University of North Carolina at Chapel Hill, LSDIs thus far have

1. Chris Anderson, "The Long Tail," *Wired* 12, no. 10 (Oct. 2004). Available online at www.wired.com/wired/archive/12.10/tail.html (Accessed 2 January 2010).
2. Lorcan Dempsey, "Reconfiguring the Library Systems Environment," *Portal: Libraries and the Academy* 8, no. 2 (2008): 111–20.

focused on general collections that can be efficiently scanned through standardized processes.[3] As we shift focus from general to special and rare collections, we need to carefully consider the lessons learned and to build on the experience gained through this first phase. The following section outlines a set of principles that need to be considered as we move the core scope of LSDIs from the general to the special and the rare.

LSDIs for Rare and Special Collections: Principles of Engagement

Develop a Business Plan to Articulate Institutional Principles

Business plans are excellent planning and communication tools for establishing a vision, allocating resources, and articulating strengths, weaknesses, opportunities, and threats (SWOT). Although often perceived only as tools for starting new businesses, they can also provide vital guidance for new initiatives, as the planning process sets an ecological vision with a thorough understanding of implications and value propositions. In their 2004 CLIR publication, Liz Bishoff and Nancy Allen provide a useful framework and resource guide to assist cultural heritage institutions with business plans for digital initiatives.[4] A business plan allows institutions considering LSDIs to identify key program issues such as rationale and objectives, market analysis, staffing, risks and benefits, resource requirements, financial information, outcomes assessment, and evaluation. Such a document is also instrumental in identifying and communicating the feasibility and desirability of entering into collaborative LSDIs and formulating a shared value proposition. For instance, Cornell University Library (CUL) identified the following principles of engagement as a foundation for its partnership with Microsoft in 2007.

Digitization Process:

- Vendor's commitment to stellar image quality and condition of the digitization facility and book handling procedures
- Quality of deliverables to CUL (master images, derivatives, metadata, etc.)
- Willingness to support an onsite facility for the digitization of rare materials
- Opportunities provided for library staff to learn from the collaboration and to influence the digitization process representing the best interests of the users and source documents

---

3. It is important to note that, although their focus has been on general collections, the Google initiative, with close to 12 million materials already digitized, also represents a significant number of rare materials.

4. Liz Bishoff and Nancy Allen, *Business Planning for Cultural Heritage Institutions* (Washington, D.C.: Council on Library and Information Resources, 2004), available online at www.clir.org/pubs/reports/pub124/contents.html (Accessed 2 January 2010).

Partnership Characteristics:

- Partners' interest in accommodating the library's interests within the context of their commercial agenda
- Nonexclusivity of contractual arrangements and ability of the library to redigitize the same materials through other initiatives
- Compatibility of the digitization project and deliverables with library's preservation plans

Access to Collections:

- Ability to provide worldwide access to the digitized collection through CUL's digital library
- Guarantees of access to digitized public domain materials with no user fees
- Ability to share the digital content with noncommercial academic initiatives and research projects
- Inclusion of rights for print-on-demand
- Inclusion of branding for the library as the materials are presented online or repurposed for print-on-demand

Set Digitization Specifications That Accommodate the Versatile Nature of Special Collections

LSDIs use a variety of digitization quality parameters that are often linked to the access requirements of the hosting companies and to the capabilities of the digitization equipment and applications. There is active debate about what is "acceptable," as well as whether current capture quality will support future viewing and processing needs. This debate is likely to get more intense as we see more representation of special collections converted through these high-throughput initiatives. Special collections are composed of rare or valuable materials and require special handling because of their scarcity, age, physical condition, monetary value, and/or security requirements. Consequently, special collections are recognized to have high digitization costs. For instance, early digitization efforts often included funds and services to prepare special and rare materials for digitization, including such activities as conservation treatment, repair or replacement of fragile pages, and rebinding. Such processes are difficult to accommodate in a large-scale initiative that favors high-speed and automated digitizing processes with uniform workflows.

A key challenge in LSDIs is balancing speed, completeness, and quality. Unlike general collections that can be digitized using standardized equipment and procedures, special and rare collections will require flexibility to assess and accommodate differing artifactual and scholarly characteristics of materials. Materials may often need to be digitized in-house to ensure secure and safe environments with curatorial oversight. Depending on acquisition arrangements, special and rare materials may

also have different security and access restrictions, such as privacy and confidentiality requirements set by donors. The 2009 Association of Research Libraries Report on Special Collections notes that, "Once archival materials are indexed in public search tools such as Google, experience has shown that they attract a steady stream of threats, takedown demands, copyright challenges, and other potential litigation."[5]

Selection decisions for LSDIs are usually made based on broad categories rather than individual assessment of titles. For instance, the Southern Historical Collection at the University of North Carolina at Chapel Hill considers its digitization large-scale because it is scanning every item of the manuscript collection with a premise that context is crucial for scholars. The archivists do not select or deselect individual documents for digitization. Another example is the Archives of American Art, an archival unit of the Smithsonian Institute, which is involved in digitizing its entire manuscript collection. Both of these institutions assume that they have already completed the selection process by acquiring these materials in their special collections. Therefore, they are comprehensively and systematically planning to digitize all material held in selected collections.[6] Such an approach necessitates that the diverse physical condition of materials is taken into consideration in selecting digitization equipment and setting digitization requirements.

Understand the Consequences of Quality Control Decisions

The scale and promise of LSDIs have prompted us to think differently about Quality Control (QC). QC is an essential component of library digitization initiatives and includes procedures and techniques to verify the quality, accuracy, and consistency of digital products encompassing images, OCR output, and other metadata files. Implementing a QC program can be very time- and labor-intensive and requires special skills and equipment. Early QC efforts made by the library community were quite thorough and often involved 100 percent QC, with visual comparison of the digital and print pages to identify subtle indicators. The Google initiative has often been criticized for producing scans with missing pages or poor image quality, and concerns have been raised about the variability of image quality and erroneous or incomplete metadata. Owing to the sheer volume of digitized content, it is not realistic to implement the kind of QC program used in past projects. As institutions convert some 10,000–40,000 books per month, it is clear that QC practices have been adjusted to suit individual budgets, technical infrastructures, staff qualifications, materials, and project timelines. Can we tolerate the same lower standards for special collection digitization initiatives? If so, what will be the consequences and risks?

---

5. Association of Research Libraries, Special Collections in ARL Libraries, *A Discussion Report from the ARL Working Group on Special Collections* (Washington, D.C.: Association of Research Libraries, 2009).

6. This approach recognizes that, inevitably, some materials will be excluded for reasons such as copyright or donor confidentiality issues.

Although we should recognize that "the perfect" may be the enemy of "the good enough," the risks taken need to be calculated with an associated contingency plan. With general collections, one of the premises is the feasibility of revisiting quality decisions at a later point and making corrections by redigitizing unacceptable sections or converting the missing elements. Although this is an expensive prospect, it is seen as an inevitable strategy that privileges the quantity of digital materials searchable on the Web over the quality of these materials. Therefore, what becomes critical is to record information about errors and omissions as they are discovered. This information is especially critical if there is full reliance on digital copies, and print versions are not readily available. Some alternative strategies for QC are presented in my 2008 white paper on large-scale digitization initiatives.[7]

Describe Collections to Support Web-Based Discovery and Access

Manual metadata creation, especially for special and rare collections, continues to be an expensive endeavor. Operating in a high-throughput environment requires that we reconsider what constitutes essential elements of metadata and how they facilitate discovery and access to materials. Metadata creation efforts need to be much more focused and move away from a "just-in-case" mode. Some institutions are experimenting with opening their content to community-based tagging so that they can be described, tagged, and commented on to leverage the knowledge of the broader community, a strategy that is being referred to as "harnessing collective intelligence." This is a worthy approach to consider, with the caveat that it be carefully designed and implemented with a robust underlying technology that facilitates the contribution process.

Finding aids will continue to be useful tools and are likely to play a stronger role in enabling discovery at a networked level. Even if the materials described in the finding aid are not available online, or are only partially digitized, the finding aids themselves are instrumental in locating collections and understanding their compositions. Although reliance on networked information that is easy to discover and access will increase, libraries and archives with special and rare collections will likely continue to draw individuals to their physical facilities. Depending on their research purposes, some scholars will persist to value interaction with material artifacts, benefiting from the visual and cognitive cues provided by the organization and presentation of special materials. Therefore, exposing the bibliographic records of holdings to search engines and union catalogs will be critical in enabling users to discover these valuable resources.

---

7. Oya Y. Rieger, *Preservation in the Age of Large-Scale Digitization: A White Paper* (2008), available online at www.clir.org/pubs/abstract/pub141abst.html (Accessed 2 January 2010).

Move from Projects to Program Mode

The mandate to move from a project mode to a program mode has almost become cliché. The principle represents the need for structuring new initiatives as embedded components of the existing organizational structure and resources to ensure their sustainability. Nevertheless, a majority of educational and cultural institutions continue to embark on digitization initiatives based solely on special funds and temporary teams, adding on rather than restructuring to streamline these new responsibilities. Structuring LSDIs as programs rather than projects includes steps such as planning the required systems for storing, archiving, and delivering the digitized content embedded in the existing technical infrastructure. The approach involves forming cross-functional teams with clear roles and leadership as well as providing staff relief where appropriate to accommodate the additional work involved in all aspects of the digitization initiative. Also essential is developing a program that creates synergy among the existing services. At Cornell, for example, we digitize content from our rare and special collections based on user requests and add the digital images to a Luna Insight collection. Materials identified by researchers and students to support their specific needs are assumed to have future users and so we scan them once, when requested, and make the digital versions available to everyone as a visual image collection.

LSDIs require substantial investments, and the scale and requirements of such initiatives put pressure on libraries and archives to move from a project to program mode. Even if digitization costs (such as materials shipping, scanning, processing, OCR creation, and indexing) are covered by commercial partners, the libraries and archives that engage in LSDIs spend significant amounts of time negotiating, planning, overseeing, selecting, creating pick lists, extracting bibliographic data, pulling and reshelving books, and receiving and managing digital content. This is an exhausting and disruptive workflow, and its associated local expenses are significant. For instance, Cornell University Library currently invests close to seven full-time equivalent staff (distributed among a total of 12 staff members) in managing LSDI-related tasks for digitizing 12,000 books per month from the general collection. But this configuration represents only the staffing required for creating pick-lists, preparing materials for digitization, reshelving materials, and supervising the production process. It does not include the resources required to ingest, archive, and provide access to the digitized materials. Often neglected or underestimated in cost analyses are the accumulated investments that libraries make in selecting, purchasing, housing, and preserving their collections.

Understand the Preservation Mandate and Requirements

The primary motivation of all partners in LSDIs is to make it easier to find and access materials. Nonetheless, access and preservation goals are usually interrelated,

since access to scholarly materials depends upon their being fit for use over time. Preservation has long been considered a fundamental responsibility of research libraries. Embarking on LSDIs prompts us to reexamine our assumptions, programs, and required resources for ensuring the longevity of digital content. The aim of the large-scale projects—to make content accessible—is interwoven with the question of how one keeps that content, whether digital or print, fit for use over time. What are the issues that will influence the availability and usability, over time, of the digital books these projects create?

The Trusted Digital Repositories Report defines *digital preservation* as "the managed activities necessary for ensuring both the long-term maintenance of a bitstream and continued accessibility of content."[8] *Bitstream preservation* aims to keep the digital objects intact and readable.[9] It ensures bitstream integrity by monitoring for corruption to data fixity and authenticity, protecting digital content from undocumented alteration, securing the data from unauthorized use, and providing media stability. At the level of enduring access, the preservation mandate also entails the processes required to provide continued access to digital content through various delivery methods. According to the *Preservation Management of Digital Material Handbook*, preserving access entails ensuring the "usability of a digital resource, retaining all quantities of authenticity, accuracy, and functionality deemed to be essential for the purposes the digital material was created and or acquired for."[10] I would like to add "preserving archival experience" as a third dimension to our overall preservation goals. This facet of archiving is elaborated in the next section.

Technology alone cannot address the goals of bitstream and enduring access preservation. Institutional policies, strategies, and funding models are also important. Although library forums began addressing digital preservation concerns almost a decade ago, only a handful of libraries today have digital preservation programs that can adequately support large-scale ingest and repository development efforts. The challenge is not only to incorporate the preservation mandate in organizational missions and programs but also to characterize the goals in a way that will make it possible to understand the terms and conditions of such a responsibility. For example, a long-term archiving mandate is likely to have different requirements than does one in support of short-term archiving goals. There are also significant differences between a preservation program that focuses on bitstream preservation

8. *Trusted Digital Repositories: Attributes and Responsibilities*, an RLG-OCLC Report (May 2002), available online at www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf (Accessed 5 January 2010).

9. "Digital objects" are items stored in a digital repository and in their simplest form consist of data, metadata, and an identifier.

10. Maggie Jones and Neil Beagrie, *Preservation Management of Digital Materials: A Handbook* (London: British Library, 2001), available online at www.dpconline.org/advice/digital-preservation-handbook.html (Accessed 10 January 2010).

and one that encompasses the processes required to provide enduring access to digital content.

As we view LSDIs from a program perspective, it is important that we embark on collaborations to collectively address the enduring access issues. On this front, we can learn from the institutional repository initiatives of the last ten years in which thousands of isolated systems were created to archive and make accessible locally produced scholarship. There is now a strong recognition that this is not a sustainable model, from both technical and service perspectives. Initiatives such as Hathi Trust, Portico, and LOCKSS are critical in addressing the collective archival needs of the educational and cultural institutions and offering solutions that are both cost-effective and sustainable.

Privilege the Users of Rare and Manuscript Collections

Digitization is not a means in and of itself but a tool to accomplish a critical mission, which is to support our users in their research, learning, and teaching activities. Context and structure are critical elements of exploring, understanding, and utilizing rare and special collections. The materiality of special collections—the physical attributes and contextual relationships—is an important cognitive element in the

research process. Users of special and rare collections often find value in their experience of physical interactions with actual materials like manuscripts and pictures. An intriguing—and potentially highly gratifying—agenda for us is to explore and understand how we can preserve the context of archival materials in digital environments. For some researchers, browsing through bookshelves or analyzing materials spread on a study desk provide the visual and spatial context that enables their critical thinking. We need to better understand the physical contextuality and spatial aspects of knowledge spaces created by our special collections. For instance, many humanists are accustomed to working in physical archives with boxes of photos, old manuscripts, diaries, and other materials. They have different colors, textures, containers. The physical world assists them in their conceptual linking.

Although today's users typically prefer to search for resources online, recent surveys and anecdotal evidence suggest that many users continue to favor a print version for reading and studying—especially for longer materials like books.[11] Some LSDI libraries are exploring the possibility of offering print-on-demand (PoD) services (especially for public domain materials) in cases where an individual contract allows it. Image quality and consistency are important factors in repurposing digitized content in support of a PoD service. Derivatives created for printing purposes have different technical requirements than do resources created to be viewed online; in the case of the former, there is heavy reliance on a high-quality master. Although the imaging requirements used by LSDIs may be "good enough" for online viewing, and even for some archival purposes, inconsistent practices and lack of quality control may impede the launch of a successful PoD program.

The popularity of search engines tends to focus our attention on research pertaining to discovery and the precision and recall of search results.[12] Although this is critical for further improvements, it is also important that we address the experience of our users after they discover materials online. Although there are usability issues that pertain to interacting with digital books online, increasing availability of diverse formats of digital content with complex relationships (such as photographs, diaries, and letters pertaining to a certain event) will further require that we pay attention to how users utilize such digital content online. Enabling discovery and access is essential in connecting our collections with users, but equally critical is ensuring that they are able to put into use the materials found, to consume and use the digital

11. According to a study at the University of Denver, most of the problems people perceive with electronic books are related to the difficulty of reading large amounts of text on the screen. The study concludes that the fact that respondents are much more likely to read portions of an electronic book than the whole could be due to that reported difficulty. Michael Levine-Clark, "Electronic Book Usage: A Survey at the University of Denver," *Libraries and the Academy* 6, no. 3 (2006): 285–99.

12. Oya Y. Rieger, "Search Engine Use Behavior of Students and Faculty: User Perceptions and Implications for Future Research," *First Monday* 14, no. 12, available online at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2716/2385 (Accessed 10 January 2010).

content according to their scholarly needs. The proliferation of viewing environments and the increasing use of mobile devices further require us to understand and to consider our users' viewing environments as we create virtual special collections.

In a report produced by OCLC Programs and Research in 2008, Merrilee Proffit and Jennifer Schaffner illustrate the critical roles rare and special materials play in both teaching and research, and they recommend specific directions for libraries and archives to address these needs.[13] Their recommendations include working with faculty to understand their research methods and materials, collaborating broadly to build collections, and continuing to build digital and material collections in support of teaching and research.

Ground Collaborations on Vigilant Negotiations

Massive digitization efforts often necessitate collaborations with commercial or nonprofit organizations such as Google or the Internet Archives. R.K. Johnson offers an excellent negotiation checklist to articulate the issues that need to be considered as an institution considers collaboration with a partner.[14] Although the list was informed by experiences with for-profit groups such as Google, several of the principles can also be used in assessing partnership opportunities with nonprofit institutions. Examples of the issues Johnson highlights include:

- What standards will be adhered to with respect to image quality, metadata, and file formats? Will specified standards be nonproprietary and sufficiently flexible to accommodate the changing environment?
- Will your institution be provided a copy of any transformation of works from your collection? Are there any restrictions on this right?
- What rights does the institution have to use the digital files? Where exclusivity may be justified, what is the maximum period over which it applies? Will restrictions lapse after a given period of time?
- Are fair use and educational protections of the copyright law preserved? Under what terms will end users have access to content?
- Will destructive processes be used in digitization? How will fragile materials be handled?
- What steps will be taken to ensure the integrity and utility of digital files?

---

13. Merrilee Proffitt and Jennifer Schaffner, *The Impact of Digitizing Special Collections on Teaching and Scholarship: Reflections on a Symposium about Digitization and the Humanities*, a report produced by OCLC Programs and Research. Available online at www.oclc.org/research/publications/library/2008/2008-04.pdf (Accessed 28 March 2010).

14. R.K. Johnson, "In Google's Broad Wake: Taking Responsibility for Shaping the Global Digital Library," *ARL: A Bimonthly Report*," Number 250, February 2007, p. 1-14, available online at www.arl.org/bm~doc/arlbr250digprinciples.pdf (Accessed 10 January 2010).

- Has the institution retained the right to make a copy of the agreement available to the public?

Concluding Remarks

The availability of scholarly content on the Web has dramatically expanded our ability to search and find scholarly materials. As the scope of LSDIs shift to special collections, it is important that we continue to assess the advances we have registered since our community embarked on digitization in the early 1990s and to explore the lessons learned to modify our strategies accordingly. As rare and manuscript collections increasingly define the uniqueness and character of individual research libraries, we have an opportunity to realign our services and strategies by taking into consideration the evolving nature of the information ecology and user expectations. What makes this task uniquely challenging is our commitment to balancing today's pressing needs with our institutional mandates and our responsibilities as stewards. Although there is an urgency and rush to digitize our cultural heritage and avail it broadly, we will succeed only if we take into consideration the long-term implications of our decisions.